# Analyses structurales et fonctionnelles de l'espace génique du chromosome 3B du blé tendre (Triticum aestivum L.)

Lise Pingault

**UNIVERSITE BLAISE PASCAL** **UNIVERSITE D'AUVERGNE**
N° D. U. 2504 ANNEE 2014

***ECOLE DOCTORALE SCIENCES DE LA VIE, SANTE, AGRONOMIE, ENVIRONNEMENT***
N° d'ordre 646

Thèse

Présentée à l'Université Blaise Pascal
pour l'obtention du grade de

**DOCTEUR D'UNIVERSITE**

(SPÉCIALITÉ : PHYSIOLOGIE ET GÉNÉTIQUE MOLÉCULAIRES)

soutenue le 30/10/2014

LISE PINGAULT

---

# Analyses structurales et fonctionnelles de l'espace génique du chromosome 3B de blé tendre (Triticum aestivum L.)

---

Président : C. TATOUT, Professeur à l'Université Blaise Pascal, Clermont-Ferrand
Membres : E. PAUX, Directeur de thèse, INRA, Clermont-Ferrand
C. RUSTENHOLZ, Maître de conférence, INRA, Colmar
Rapporteurs : B. CHALHOUB, Directeur de Recherches, INRA, EVRY
S. ROBIN, Directeur de Recherches, INRA, Paris
M. TENAILLON, Directrice de Recherches, CNRS, Gif-sur Yvette

UMR 1095 INRA-UBP « Génétique, Diversité et Ecophysiologie des Céréales »
5 Chemin de Beaulieu, 63039 Clermont-Ferrand Cedex 2

**Résumé**

De par sa taille (17 Gb), la complexité de son génome (allohexaploïde) ainsi que la forte proportion d'éléments répétés (>80%), l'étude du génome de blé tendre est une tâche particulièrement complexe et s'est souvent retrouvée confrontée aux limites technologies. Grâce une approche de tri de chromosomes, le chromosome 3B (995 Mb) a pu être isolé et séquencé. Ces données ont permis la construction d'une pseudomolécule. Mes travaux de thèse se sont basés sur des données de transcriptomique produites avec une approche RNA-Seq, afin d'investiguer l'impact de la taille de ce chromosome sur l'organisation de l'espace génique.

L'annotation du chromosome 3B a permis de mettre en évidence : 5 326 gènes et 1 938 pseudogènes. L'analyse des librairies RNA-Seq pour 15 conditions de développement a permis de mettre en évidence l'expression de 71 % des gènes annotés, ainsi que 3 692 régions nouvellement transcrites (NTR). Nous avons aussi pu détecter des transcrits alternatifs pour 61% des gènes exprimés (en moyenne 6 isoformes).Nous avons donc pu mettre en évidence une structuration de l'espace génique pour le chromosome 3B. En effet, la transcription est répartie sur tout le chromosome, cependant les gènes sont organisés selon un gradient de densité croissant sur l'axe centromère-télomère. En nous basant sur le profil des données de recombinaison, nous avons divisé le chromosome en 3 régions : R1, R2 et R3. La région R2 correspondant à la région centrale du chromosome (647 Mb) où le taux de recombinaison est très faible voir absent. Les régions R1 (58 Mb) et R3 (69 Mb) correspondent respectivement aux parties distales du bras court et du bras long du chromosome, où le taux de recombinaison est le plus fort. Ces trois régions diffèrent par leur niveau et leur spécificité d'expression, ainsi que par leur structure génique (nombre d'exons, taille des introns …). En effet, les gènes ayant une expression tissu-spécifique, ainsi qu'un faible nombre de transcrits alternatifs sont retrouvés dans les régions R1 et R3. Deux modèles peuvent expliquer le lien observé entre la structure des gènes et leur niveau/spécificité d'expression : le modèle de la sélection pour l'économie et le modèle dessin génomique.

En conclusion, ce travail a montré et ce, pour la première fois à l'échelle d'un chromosome entier de blé, l'impact de la taille du chromosome sur l'organisation; mettant en relation la structure des gènes, leur niveau d'expression, leur spécificité d'expression, ainsi que leur nature évolutive. L'assemblage ainsi que l'annotation de pseudomolécules des autres chromosomes permettra de mettre en évidence si cette structure est conservée. Afin de mieux comprendre les mécanismes cellulaires impliqués dans la régulation de l'expression des gènes, une étude du paysage épigénomique a été engagée.

Mots clés : blé, espace génique, RNA-Seq, expression, structure du génome.

**Abstract**

Genome-wide studies of the bread wheat are a complicated task due to its large size (17 Gb), its allohexaploidy and its high content in repeat sequences (>80%). Using a chromosome-specific approach, the chromosome 3B (995 Mb) was successfully isolated and sequenced leading to the assembly of one pseudomolecule. The work presented in this thesis investigated the impact of the 3B chromosome size on the gene space organization. Production of transcriptomic data was achieved using RNA-Seq approach.

The chromosome 3B was annotated and we predicted 7 264 features, including 5 326 full genes and 1 938 pseudogenes. We constructed RNA-Seq libraries for 15 developmental wheat conditions. Using this data we detected expression of 71.4% of the predictions, and 3 692 novel transcribed regions (NTR). We also detected alternative transcripts for 61% of the expressed genes, with 5.8 isoforms on average for one gene. Using these transcriptional data, we highlighted a partitioning of the chromosome 3B gene space. Indeed, transcription was found all along the chromosome, but genes were organized according to an increasing density gradient along the centromere-telomere axis. Based on recombination profile, we segmented the chromosome in 3 major regions: R1, R2 and R3. The region R2 was identified with low or no recombination rate corresponding to the centromeric and peri-centromeric regions (647 Mb). The regions R1 and R3 were associated with a higher recombination rate, both localized on the distal part of the short arm (58 Mb) and the long arm (69 Mb) respectively, where the recombination rate is higher. All three regions showed distinct level and specificity of gene expression as well as unique gene structure (variation size, exon number, intron size). Indeed, genes expressed in a specific condition and with a small number of alternatives transcripts were localized on regions R1 and R3. We showed that two evolutionary model could explain the link between gene structure and the level/specificity of expression : “selection for economy” and “genome design”.

In conclusion, a transcriptomic studies was achieved along the 3B chromosome for the first time. This study demonstrated a relationship between gene characteristics (structure, expression level, expression specificity and evolution) and the chromosome 3B organization. Future pseudomolecule assemblies will help us to assess the structural organization of these chromosomes. In order to better understand the cellular mechanisms of gene expression, an epigenomic study of the 3B chromosome was started.

Key words: wheat, gene space, RNA-Seq, expression, genome structure.

# Remerciements

Je tiens avant tout à remercier mes parents, premiers sponsor et supporters de cette thèse.

Je tiens aussi à remercier les deux directeurs d'unité qui se sont succédés au cours de ma thèse : Gilles Charmet et Thierry Langin, pour m'avoir accueilli dans l'unité Génétique, Diversité et Ecophysiologie des Céréales.

Je remercie aussi la Région Auvergne et le FEDER, pour le financement de la thèse.

Un grand merci à Catherine Feuillet, pour sa disponibilité, ses connaissances, son enthousiasme et sa simplicité.

Je remercie aussi Etienne Paux (The Boss), pour m'avoir donné la possibilité de participer à ce beau et grand projet. Merci aussi pour tes conseils et les discussions qui m'ont permis d'élaborer ce manuscrit.

Fred Choulet, merci pour tes précieux conseils, ta pédagogie, ta patience, ton implication dans ce manuscrit.

Je tiens aussi à remercier tous les membres de l'équipe Génome (ancienne et nouvelle) pour leur accueil et leur bonne humeur.


Je remercie plus particulièrement mes voisines/voisins de bureau qui se sont succédés pendant plus de 3 ans : Christel, Hélène, Aurélie, Julien, pour leur aide, les bons moments de délire …

Je remercie aussi l'équipe café, pour les bons moments de rigolade ^^ et les bons gâteaux (surtout les muffins d'Audrey)

Merci à Aurélien B. pour le « financial support » ;)

Merci à la team plateforme (Charles, Véronique, Géraldine, Karine, Lydia, Anthony) pour leur Shaka attitude, et aussi pour m'avoir trouvé si gentiment une petite place pendant près de 2 mois.

Je remercie l'ensemble du personnel de l'unité GDEC. Ainsi que toutes les personnes qui ont été impliquées de près ou de loin dans ce projet.

# Sommaire

## Avant-propos

Aujourd'hui, le blé tendre (*Triticum aestivum*) est la céréale la plus cultivée dans le monde avec 215 millions d'hectares, et représente la nourriture de base pour 30% de la population mondiale. C'est la deuxième source majeur d'apport caloriques après le riz (532 kcal par personne et par jour, soit 20% des apports caloriques) (This week in Nature - Editorials, 2014).

Depuis 1960, les rendements du blé ont triplé, passant de 1,1 t/Ha (1961) à 3,1 t/Ha (2012) (« FAOSTAT », 2014). Cette augmentation s'inscrit dans la période de la « Révolution Verte », qui a permis d'améliorer et de produire des variétés de blé par le biais de programmes de recherche innovants, ainsi que par l'application de nouvelles pratiques culturales (irrigation, utilisation d'engrais azotés) (P. S. Baenziger, Russell, Graef, & Campbell, 2006). Cependant, sur les dix dernières années, le rendement du blé a stagné avec une augmentation d'environ 0,9% par an (This week in Nature - Editorials, 2014). Ce qui contraste avec l'augmentation de la consommation, due aux changements d'habitudes alimentaires (augmentation de la consommation de la viande estimée à 70% d'ici 2050), ainsi qu'à l'augmentation de la taille de la population mondiale (en 2050, la taille de la population est estimée à 9,6 milliards, soit 2,4 milliards de consommateurs supplémentaires). Il faut aussi prendre en compte la taille des surfaces agricoles qui diminue, avec l'urbanisation, la concurrence des surfaces de pâturage ou bien de bois de chauffe, ainsi que la production de biocarburant (Rice & Garcia, 2011).

A la vue des ces nouvelles données, l'agriculture mondiale se trouve face à un défi sans précédant : (i) répondre à l'augmentation démographique ainsi qu'à de nouveaux besoins alimentaires, (ii) cela dans un cadre d'une agriculture durable et responsable. La FAO estime qu'il faudrait augmenter la production de blé de 2,4% par an, pour atteindre une augmentation totale de 60% d'ici 2050. Tout en respectant une politique agricole, qui vise une agriculture plus propre, par la réduction de l'utilisation des produits phytosanitaires. De plus, les changements climatiques sont de plus en plus importants et impactent sur les rendements.

On constate que la stagnation des rendements observée pour le blé n'est pas retrouvée chez les autres céréales, dont le maïs en particulier. Pour cela trois raisons principales : (i) le fort investissement en recherche et développement, qui chez le maïs est quatre fois supérieur à celui fait sur le blé (This week in Nature - Editorials, 2014), (ii) l'adoption rapide des nouvelles technologies pour les programmes d'amélioration variétale, (iii) l'utilisation de

semences hybrides (Whitford et al., 2013). Un point clé impliqué dans ces trois raisons : l'absence d'une séquence de référence pour le blé hexaploïde.

En effet, une grande partie des génomes des céréales les plus cultivées ont déjà été séquencés, comme par exemple : le maïs, le riz ou bien le sorgho. Cependant, malgré son importance socio-économique, le blé résiste quant au décryptage de son génome, du fait de sa taille et de sa complexité. Ce qui fait de son génome un excellent modèle pour les espèces à génome complexe.

C'est dans ce cadre que s'inscrit mon sujet de thèse, en proposant pour la première fois une analyse complète de l'organisation structurale et fonctionnelle de l'espace génique pour un chromosome de blé.

# Glossaire

| | |
|---|---|
| ADN | : Acide Désoxyribonucliéique |
| ADNc | : ADN complémentaire |
| ADNg | : ADN génomique |
| ARN | : Acide Ribonucléique |
| ARNnc | : ARN non codant |
| ARNr | : ARN ribosimique |
| ARNt | : ARN de transfert |
| BAC | : Bacterial Aritifical Chromosome |
| CDS | : Coding Sequence |
| CGH | : Comparative Genomic Hybridization |
| CNV | : Copy Number Variation |
| CS | : Chinese Spring |
| EST | : Expressed Sequenced Tag |
| ET | : Elément Transposable |
| FAO | : Food and Agriculture Organization of the United Nation |
| Gb | : Giga base |
| GO | : Gene Ontology |
| IWGSC | : International Wheat Genome Sequencing Consortium |
| kb | : kilo base |
| lincRNA | : long intergenic non coding RNA |
| lncRNA | : long non coding RNA |
| Mb | : Méga base |
| mC | : méthylation de la Cytosine |
| miRNA | : micro RNA |
| MPT | : Modification Post-Transcriptionnelle |
| MTP | : Minimum Tilling Path |
| NGS | : Next Generation Sequencing |
| nt | : nucléotide |
| ORF | : Open Reading Frame |
| PAV | : Presence Absence of Variation |
| pb | : paire de base |
| PCR | : Polymerase Chain Reaction |
| pg | : picogramme |
| PSE | : Promoteur Spécifique de l'Expression |
| RNA-Seq | : RNA Sequencing |
| SAGE | : Serial Analysis of Gene Expression |
| siRNA | : small interfering RNA |
| SNP | : Single Nucleotide Polymorphism |
| TSS | : Transcription Start Site |
| WCS | : Whole Chromosome Shotgun |
| WGD | : Whole Genome Duplication |
| WGS | : Whole Genome Shotgun |
| YAC | : Yeast Artificial Chromosome |

## Table des Figures

# Table des Tableaux

# INTRODUCTION

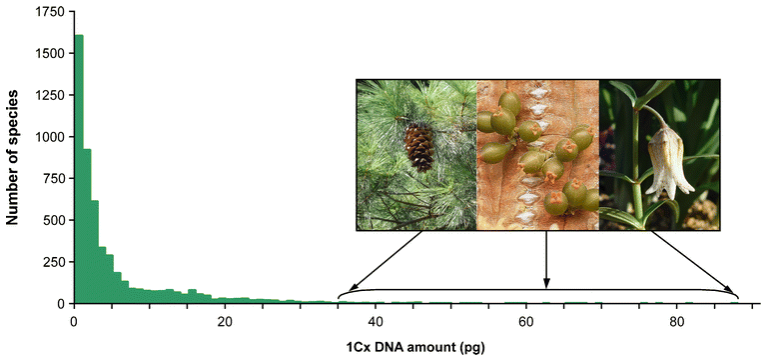



**Figure 1** : Histogramme représentant la taille de génomes monoploïdes.
La taille est calculée par la division de valeur de 2C par le niveau de ploïdie pour 5173 espèces. Les valeurs ont été représentée par tranches de 1 pg (allant de 1 pg à 87-88 pg). L'accolade représente les espèces qui ont des génomes de grande taille (1C >35 pg). Les trois espèces représentées sont (de gauche à droite): Pinus ayacahuite (gymnosperme), Viscum minimum (eudicotylédone), Fritillaria japonica (monocotylédone).(Kelly et Leitch 2011).

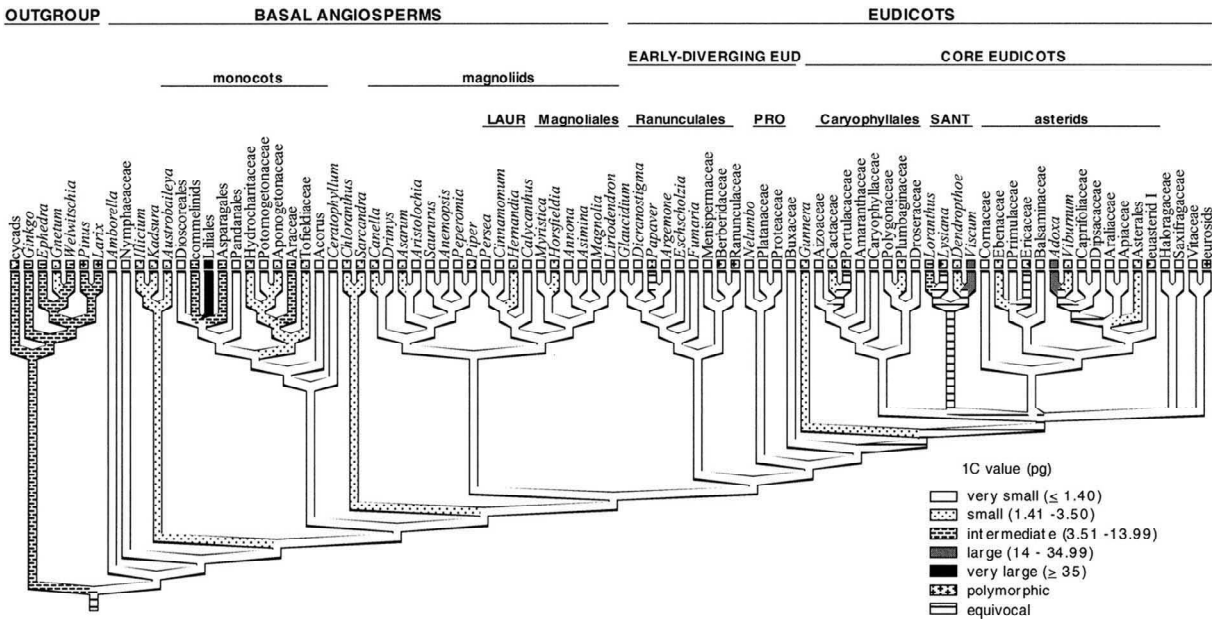



**Figure 2** : Représentation de la diversité de la taille des génomes chez les Angiospermes.
La méthode de construction de l'arbre utilisée est la parcimonie. D'après (Soltis et al. 2003).

# 1 Structure et organisation des génomes végétaux

## 1.1 Les variations de taille des génomes chez les plantes

La taille des génomes de plantes varie de quelques Mb (ex : 63 Mb chez le genre *Genlisea* (Greilhuber et al., 2006)) à plusieurs dizaines de Gb (ex : *Paris japonica* : 148 852 Mb (Pellicer, Fay, & Leitch, 2010)). Traditionnellement, la taille d'un génome est rapportée à sa « valeur C » qui représente le contenu en ADN pour une cellule haploïde. Elle est mesurée en picogrammes (pg) avec la relation suivante : 1 pg représente 978 Mb (http://www.genomesize.com/). A partir des valeurs issues de la base de données « Plant DNA C-values data base » (http://data.kew.org/cvalues/), qui recense les valeurs C connues des génomes de plantes, on constate que la valeur C varie d'un facteur 80 (Figure 1). Toutefois, cette valeur n'est pas corrélée au nombre de gènes codant une protéine. Ce phénomène est appelé le « paradoxe de la valeur C » (Pagel & Johnstone, 1992). L'explication réside dans le fait que la taille du génome nucléaire est fonction de la fraction non génique du génome. En effet, il a été mis en évidence que la rapidité du changement de taille des génomes est due à l'activité des éléments transposables (ET) (Tenaillon, Hollister, & Gaut, 2010).

Les génomes de plantes ont été classés en cinq groupes en fonction de leur taille : très petits, petits, moyens, grands et très grands, à partir de l'estimation des valeurs C pour 1% des génomes de plantes (I. J. Leitch, Soltis, Soltis, & Bennett, 2005). Les très petits génomes et les petits génomes sont ceux avec une valeur C inférieure ou égale à 1,4 pg et inférieure ou égale à 3,5 pg respectivement, les génomes de taille moyenne sont compris entre 3,5 et 14 pg, les génomes de grande taille ont une valeur C supérieure ou égale à 14 pg et inférieure à 35 pg. Enfin, les génomes avec une valeur C supérieure ou égale à 35 pg sont considérés comme des génomes de très grande taille (I. J. Leitch et al., 2005). La majorité des plantes dont la taille du génome a été estimée ont un petit génome (Kelly & Leitch, 2011). Cependant, on retrouve des plantes à grand génome dans plusieurs clades, les Liliales (monocotylédones), les Santalales (dicotylédones), et les Ophioglossales (fougères), ce qui indique que l'expansion de la taille des génomes se serait produite indépendamment dans plusieurs lignées au cours de l'histoire évolutive (I. J. Leitch et al., 2005).

Chez les Angiospermes (division des *Magnoliophyta*), la taille des génomes est fortement variable (Figure 2) (Soltis, Soltis, Bennett, & Leitch, 2003). La moitié des espèces ont un génome de taille inférieure ou égale à 3,5 pg (3 423 Mb). En revanche, les génomes de

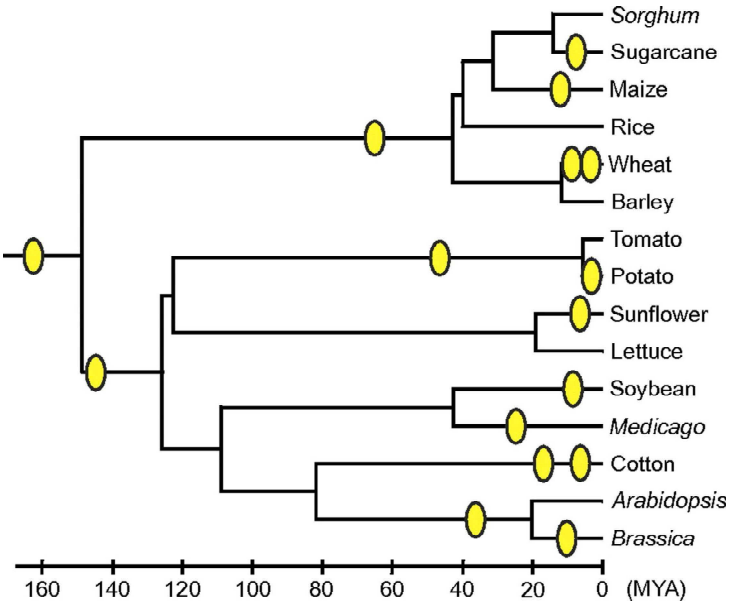



**Figure 3** : Représentation des différents évènements de polyploïdisation durant l'évolution des Angiospermes.
Les cercles jaunes indiquent les évènements de duplication. L'échelle du temps est donnée en million d'années (Nei et Nozawa 2011).

grande taille (C ≥ 14 pg) ont une distribution beaucoup plus réduite, et sont retrouvés uniquement chez quelques monocotylédones, Ranunculades, Caryophyllales et Santales (Soltis et al., 2003). L'étude des génomes complexes permet l'étude des mécanismes moléculaires et des forces évolutives favorisant ou contrecarrant l'expansion des génomes. Elle permet ainsi de mieux comprendre comment et pourquoi la plupart des génomes sont contraints en taille (Kelly & Leitch, 2011).

## 1.2 Origine des variations de tailles : polyploïdisation et éléments transposables

### 1.2.1 Polyploïdisation

La polyploïdie était initialement considérée comme « une entrave au succès de l'évolution des plantes » (Stebbins, 1971). Aujourd'hui, elle est considérée comme une force majeure de l'évolution chez les plantes, affectant la taille des génomes, mais surtout la diversification et la spéciation (Adams & Wendel, 2005; Madlung, 2013; S P Otto & Whitton, 2000; Soltis et al., 2009) (Figure 3). Chez les espèces à génome polyploïde, le caryotype n'est pas composé de paires de chromosomes homologues (diploïde 2x) mais de plusieurs exemplaires de chacun des chromosomes (exemples : banane : 3x ; canne à sucre : 8x). Cette redondance permet la divergence moléculaire et fonctionnelle d'une des copies de gènes dupliqués, favorisant ainsi la plasticité du génome (Bento, Gustafson, Viegas, & Silva, 2011). Ce phénomène est beaucoup plus fréquent chez les plantes que chez les animaux. En effet, chez les plantes à fleurs, ce phénomène se produit à une fréquence de 1 sur 100 000 (Comai, 2005) alors que sa prévalence est beaucoup plus faible chez les animaux avec seulement 200 exemples de polyploïdies rapportés chez les insectes et les vertébrés (S P Otto & Whitton, 2000). Chez les mammifères et les oiseaux, les changements d'état de ploïdie peuvent être fatales, notamment à l'état embryonnaire (exemple : chromosomes sexuels sur-numéraires) (Sarah P Otto, 2007).

Chez les plantes à fleurs, plus de 70% des espèces sont des polyploïdes (Graham Moore, 2002), à la différence des gymnospermes où un faible pourcentage des espèces sont polyploïdes (A. R. Leitch & Leitch, 2012). De nombreuses espèces végétales cultivées ont un génome polyploïde, comme le blé, le café, la banane, la pomme de terre, la canne à sucre, le tabac (A. R. Leitch & Leitch, 2008). L'événement de polyploïdisation est, généralement, rapidement suivi de modifications épigénétiques, de la perte massive de copies de gènes dupliqués et de réarrangements structuraux, dues au fait que la redondance ainsi acquise qui ne va pas être sélectionnée (Comai, 2005).

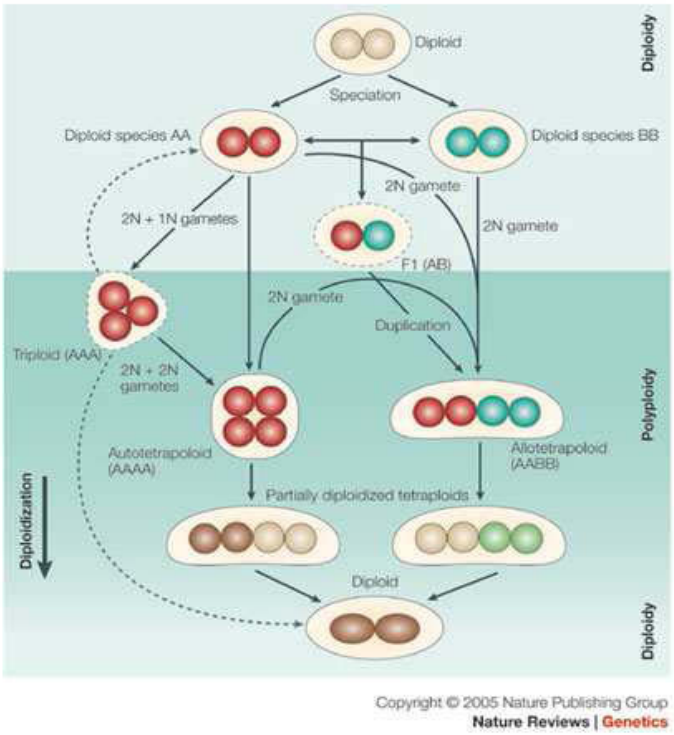



**Figure 4** : Schéma de l'alternance évolutive de l'état de ploïdie d'une cellule.
Pour chaque forme de ploïdie, le génome haploïdie est représenté par une forme circulaire ou ovale colorée, à l'intérieur d'une cellule (fond beige). Les génomes de forme ovale montrent l'augmentation du nombre de gènes qui résultent d'une rétention et d'une sous fonctionnalisation durant la diploïdisation. Les formes de ploïdies hautement instables sont représentées avec des contours en pointillés. Les lettres A et B représentent deux types de génomes différents, et N définit le nombre de chromosomes gamétiques (Comai 2005).

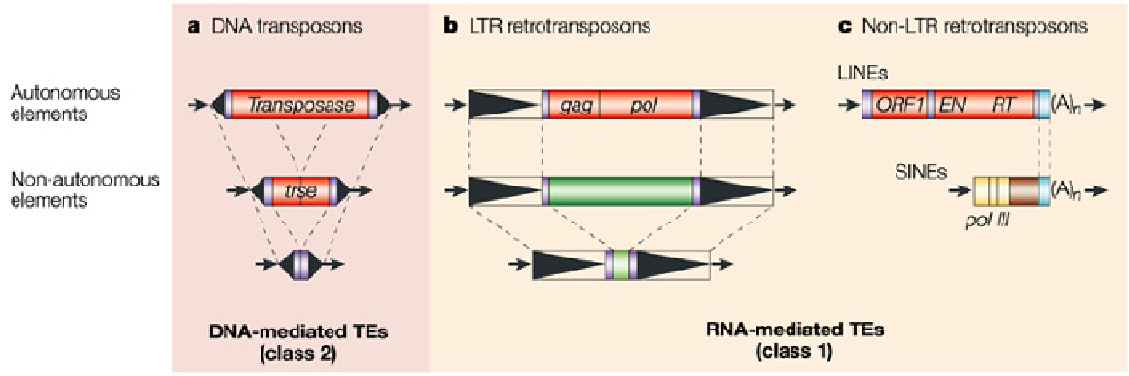



**Figure 5** : Classification et structure des éléments transposable chez les Eucaryotes.
(a) les éléments de classe 2, (b et c) les éléments de classe 1. Les éléments de classe 1 sont divisés en deux groupes en fonction de leur mode de transposition et de leur structure : les retrotransposons à LTR (b) et les rétrotransposons sans LTR (c). Dans chaque classe, ont retrouve des éléments autonomes et non-autonomes. Les éléments autonomes codent pour des protéines requises pour la transposition (gag, capside like-protein; pol, transcriptase inverse; ORF1, une protéine gag-like; en, endonucléase; rt, transcriptase inverse). Alors que les éléments non-autonomes ne codent pas pour ces protéines mais captent des séquences cis, nécessaires à la transposition (Wessler 2006).

Il existe deux types de polyploïdie : autopolyploïdie et allopolyploïdie. L'autopolyploïdie se réfère à la multiplication de chromosomes génétiquement identiques d'une espèce, et résulte d'anomalies méiotiques ou d'hybridation intra spécifique, alors que l'allopolyploïdie entraîne la multiplication d'un lot de chromosomes par l'hybridation de deux espèces différentes (Comai, 2005; Weiss-Schneeweiss, Emadzade, Jang, & Schneeweiss, 2013). Dans ce cas, les sous-génomes issus des deux parents sont qualifiés d'homéologues (S P Otto & Whitton, 2000) (Figure 4).

Trois avantages majeurs de l'état polyploïde ont été recensés (Comai, 2005). Les deux premiers : la vigueur hybride et la redondance des gènes, sont liés à la duplication des gènes. Le troisième est la reproduction asexuée, afin de perturber certains systèmes d'auto-incompatibilité. Des désavantages ont aussi été décrits, comme par exemple les effets perturbateurs de l'élargissement nucléaire et cellulaire, ainsi que les problèmes rencontrés en mitose et méiose, entrainant la formation de cellules aneuploïdes (Comai, 2005).

### 1.2.2 Les éléments transposables

Si la polyploïdie joue un rôle majeur dans l'évolution de la structure et de la taille des génomes chez les plantes, les éléments transposables sont les principaux acteurs de l'expansion des génomes, d'où le « paradoxe de la valeur C » énoncé précédemment (Kidwell, 2002). Les éléments transposables ont été découverts dans le milieu des années 1940 par Barbara McClintock chez le maïs. Ils représentent plus de 70% de certains génomes végétaux et sont mobiles dans un génome (Wessler, 2006). Chez les eucaryotes, les ET sont divisés en deux classes en fonction de leur mode de transposition. Les ET dits de classe I, sont ceux qui transposent via un intermédiaire ARN, alors que les ET qui transposent via un intermédiaire ADN sont qualifiés de ET de classe II. Dans chaque classe, il existe des éléments autonomes (qui codent pour des protéines nécessaires à la transposition) et non autonomes qui ont besoin d'un partenaire autonome au sein du même génome pour être mobilisés. Chez les éléments de classe I, on distingue les rétrotransposons à LTR (gypsy, copia) et ceux sans LTR (LINE, SINE). Les rétrotransposons à LTR représentent généralement les ET les plus abondants dans les génomes de plantes (Bennetzen & Wang, 2014; Wessler, 2006) (Figure 5). Chez les Angiospermes, les génomes de grande taille dérivent de l'amplification massive d'un petit nombre de familles de rétrotransposons à LTR (Bennetzen & Wang, 2014). Ceux-ci auraient tendance à s'insérer dans les rétrotransposons à LTR déjà présents (Bennetzen, 2002). A la différence de la compaction du génome, qui se produit par délétion de fragments d'ADN via, notamment, la recombinaison homologue inégale entre paires de LTR (d'un même élément ou entre deux

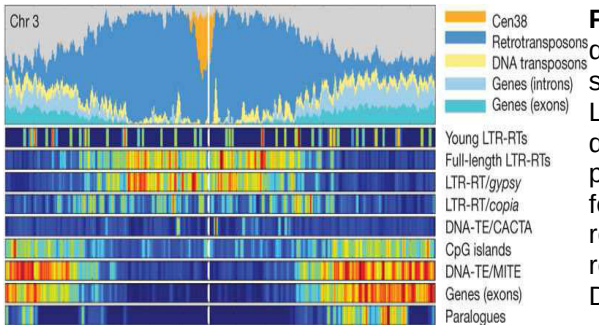
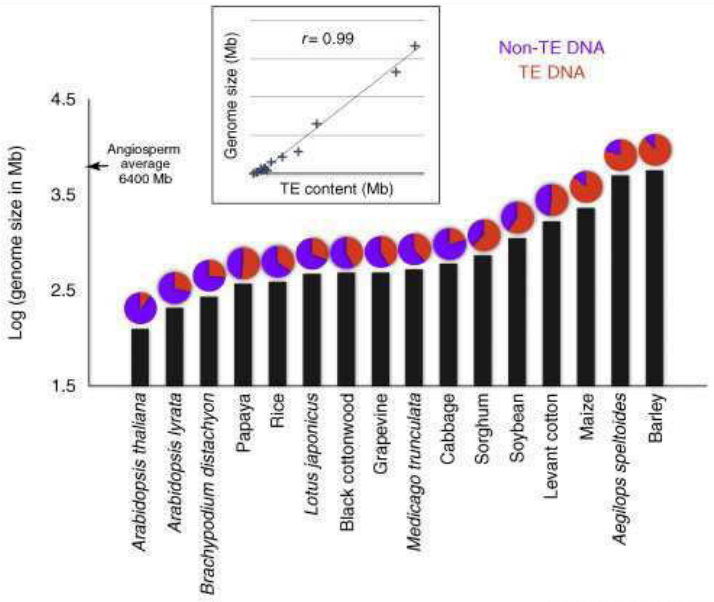



**Figure 6** : Représentation de l'organisation des différentes familles de TE et des gènes chez le sorgho.
Les « Heat map » représentent les variations des différents éléments. Le graphique représente les proportions (en pb) de chaque élément sur une fenêtre glissante de 0,5 Mb. Le gris indique les régions non assignées ainsi que les régions régulatrices.
D'après (Paterson et al. 2009).

**Figure 7** : Représentation de la proportion de TE dans des génomes de différentes taille.
La taille des génomes est fortement corrélée avec de contenu en ET chez les Angiospermes (Tenaillon, Hollister, et Gaut 2010)




| Nom de l'espèce | Nom commun | Taille (Mb) | Chromosomes | Ploïdie | # gènes | % ET | Stratégie de séquençage |
|---|---|---|---|---|---|---|---|
| Actinidia chinensis | Kiwi | 616,1 | 29 | diploïde | 39 040 | 36% | WGS |
| Arabidopsis lyrata | Arabette Lyrates | 207 | 8 | diploïde | 32 670 | 29,7% | BAC |
| Arabidopsis thaliana | Arabidopsis | 125 | 5 | diploïde | 33 602 | 14% | BAC |
| Brachypodium distachyon | Faux brome violet | 272 | 5 | diploïde | 32 255 | 21,4% | BAC |
| Brassica rapa | Choux chinois | 485 | 10 | diploïde | 22 285 | 39,5% | BAC/WGS |
| Cicer arietinum | Pois Chiche | 738 | 8 | diploïde | 28 269 | 32,2% | WGS |
| Carica papaya | Papaye | 372 | 9 | diploïde | 24 746 | 51,9% | WGS |
| Chlamydomonas reinhardtii | Algue verte | 121 | 17 | diploïde | 16 036 | | shotgun/BAC |
| Cucumis sativus | Concombre | 243,5 | 7 | diploïde | 26 682 | 14,8% | WGS |
| Fragaria vesca | Fraise | 240 | 7 | diploïde | 34 809 | 22% | WGS |
| Glycine max | Soja | 1100 | 20 | diploïde | 66 153 | 59% | WGS |
| Gossypium raimondii | Coton | 737,8 | 13 | diploïde | 37 505 | 61% | BAC |
| Lotus japonicus | Lotus | 472 | 6 | diploïde | 42 399 | 34,28% | BAC/WGS |
| Musa acuminata | Banabe | 523 | 11 | haploïde doublé | 36 542 | ~50% | WGS |
| Malus x domestica | Pomme | 742 | 17 | diploïde | 57 386 | 42,4% | WGS |
| Medicago truncatula | Luzerne tronquée | 375 | 8 | diploïde | 45 108 | | BAC/WGS |
| Oryza sativa | Riz | 389 | 12 | diploïde | 35 679 | 35% | BAC |
| Populus trichocarpa | Peuplier | 485 | 19 | diploïde | 45 778 | 42% | |
| Sorghum bicolor | Sorgho | 730 | 10 | diploïde | 34 496 | 61% | WGS |
| Solanum lycopersicum | Tomate | 900 | 12 | diploïde | 34 727 | 63,2% | BAC/WGS |
| Solanum tuberosum | Pomme de terre | 844 | 12 | diploïde | 39 031 | 62,2% | WGS |
| Theobroma cacao | Cacao | 430 | 10 | diploïde | 28 798 | 25,7% | WGS |
| Vitis vinifera | Vigne | 487 | 19 | diploïde | 30 434 | 41,4% | WGS |
| Zea mays | Maïs | 2300 | 10 | diploïde | 32 540 | 85% | BAC |

**Tableau 1** : Tableau récapitulatif non exhaustif des génomes de plantes séquencés.
D'après http://chibba.agtec.uga.edu/duplication/

éléments de la même famille) et donne naissance aux solo-LTR (Devos, Brown, & Bennetzen, 2002). La variation de la taille des génomes de plante est donc due à la balance entre accumulation et perte des ET.

La distribution de certaines familles d'ET le long des chromosomes peut être très hétérogène avec une tendance générale à l'accumulation des ET dans les régions centromériques comme décrit chez le sorgho et *A. thalinana* (Arabidopsis Genome Initiative, 2000; Paterson et al., 2009a) (Figure 6). Ainsi, si le contenu en gène reste similaire entre les espèces, le pourcentage de ET varie significativement et une corrélation forte existe entre la taille des génomes de plantes et leur contenu en LTR-rétrotransposons (Tenaillon et al., 2010) (Figure 7).

## 1.3 La distribution des gènes dans l'espace génique

### 1.3.1 Variation de la densité de gènes

L'espace génique est définit comme la distribution des gènes le long des chromosomes (Barakat, Carels, & Bernardi, 1997). En 2004, l'espace génique est redéfinit en référence à la portion de l'ADN codant des gènes actifs et ainsi qu'à la distribution des gènes actifs dans un génome (Scott Jackson, Barbara Hass Jacobus, & Janice Pagel, 2004). Plus généralement, l'espace génique se rapporte à de grandes régions riches en gènes, séparées par de grandes régions pauvres en gènes (Varshney, Hoisington, & Tyagi, 2006).

Chez les plantes, le contenu en gènes est relativement constant : entre 16 036 chez *Chlamydomonas reinhardtii* (algue verte ; (Merchant et al., 2007)) et 66 153 chez *Glycine max* (soja ; (Schmutz et al., 2010)), soit une variation d'un facteur de 4,12 (Lee, Tang, Wang, & Paterson, 2012) alors que leur génome représentent respectivement 121 Mb et 1,1 Gb (Tableau 1).

Par ailleurs, l'ordre des gènes (ou colinéarité) est conservé entre les génomes apparentés (G Moore, Devos, Wang, & Gale, 1995). Toutefois, les gènes ne sont pas distribués aléatoirement et une structuration de l'espace génique a été décrite en fonction de la taille des génomes (D'Hont et al., 2012; International Brachypodium Initiative, 2010; International Rice Genome Sequencing Project, 2005; Paterson et al., 2009a; Schnable et al., 2009). Pour les génomes de taille inférieure à 500 Mb environ, la distribution des gènes a tendance à être uniforme. C'est le cas pour les génomes de *C. reinhardtii* (121 Mb; (Merchant et al., 2007), *A. thaliana* (125 Mb; (Arabidopsis Genome Initiative, 2000), *B. distachyon* (272 Mb ; (International Brachypodium Initiative, 2010)) et du riz (389 Mb ; (International Rice Genome Sequencing Project, 2005)). En revanche, chez le maïs (2,3 Gb ; (Schnable et al., 2009)), le

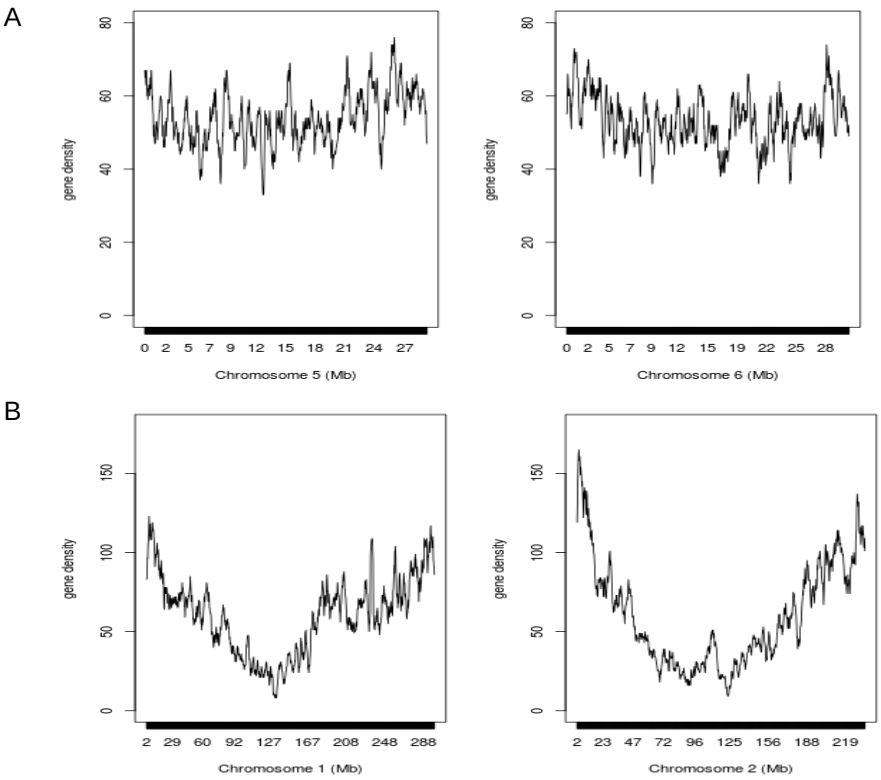



**Figure 8** : Densité de gènes le long des chromosomes 5 et 6 du riz, et 1 et 2 du maïs.
(A) densité de gènes pour les chromosomes 5 et 6 du riz, sur une fenêtre glissante de 0,35 Mb. D'après (http://rice.plantbiology.msu.edu/)
(B) densité de gènes pour les chromosomes 1 et 2 du maïs, sur une fenêtre glissante de 3 Mb. D'après (http://ensembl.gramene.org/Zea_mays/Info/Index)

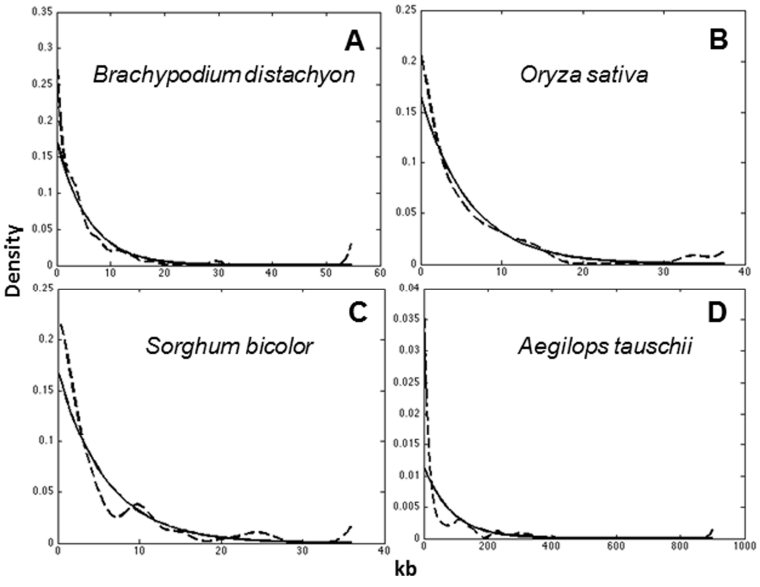



**Figure 9** : Illustration de la fonction de densité des distances intergéniques.
Pour les quatre génomes de plante, le trait plein représente la densité exponentielle qui est rapportée au maximum de vraisemblance, et en pointillé est représentée l'estimation de la densité non paramétrique. D'après (Gottlieb et al. 2013).

sorgho (730 Mb ; (Paterson et al., 2009a)), ou encore l'orge (5,1 Gb ; (International Barley Genome Sequencing Consortium et al., 2012)), une augmentation de la densité de gènes dans les parties distales des chromosomes a été mise en évidence. Par exemple la densité en gènes des chromosomes de riz varie entre 14 et 16 gènes par Mb alors qu'elle varie de 0,5 à 5 gènes par Mb chez le maïs (Figure 8).

### 1.3.2 L'organisation en insulas et gènes co-exprimés

Au delà du gradient de densité en gènes observé le long des chromosomes chez les génomes de grande taille, à une échelle plus fine, une tendance au regroupement des gènes en îlots a été décrit chez plusieurs espèces (Par exemple : le blé : Devos et al. 2005; Choulet et al. 2010; Rustenholz et al. 2011 ; le coton : (W. Guo et al., 2008)). Chez le coton, le critère choisi pour définir des îlots de gènes est une distance intergénique inférieure à 5 kb. Chez le blé, c'est la valeur médiane des distances intergéniques qui a été choisie comme seuil. Plus récemment, une approche statistique pour définir les îlots a été proposée (Gottlieb et al., 2013). En se basant sur un échantillon de séquences de BAC d'*Ae. tauschii*, les auteurs ont définit l'hypothèse nulle suivante : les gènes sont répartis aléatoirement sur les chromosomes suivant une distribution uniforme, apparentée à une loi de Poisson, supposant que les gènes sont localisés indépendamment les uns des autres. Le pendant de cette hypothèse est que la densité des distances intergéniques suit une distribution exponentielle. Cette étude a montré que pour les génomes compacts, comme ceux du riz et de *B. distachyon*, il n'y a pas de différence significative entre les densités de gènes observées et celles proposées par le model exponentiel, alors que cette différence est significative pour les génomes plus grands du sorgho et d'*Ae. tauschii* (Figure 9). La dénomination « d'insulae » (insula) a ainsi été proposée pour décrire un cluster de gènes proches.

Précédemment, nous avons montré que la taille des génomes était fortement impactée par les ET. En effet, les génomes qui ont peu de régions denses en gènes ont un faible pourcentage de rétroéléments, comme celui du riz (35% ; (International Rice Genome Sequencing Project, 2005)) ou de *B. distachyon* (21.4% ; (International Brachypodium Initiative, 2010)), à la différence de celui du sorgho (61% ; (Paterson et al., 2009b)), du maïs (>75% ; (Schnable et al., 2009)) où d'*Ae. tauschii* (65,9% ; (Jia et al., 2013)).

Les gènes retrouvés en insula sont généralement aussi retrouvés co-exprimés. Chez *A. thaliana* 10% des gènes sont co-exprimés, avec une partie important sont des gènes dupliqués (Zhan, Horrocks, & Lukens, 2006). Cependant la duplication seule des gènes ne suffit pas à expliquer le pourcentage de gènes retrouvés co-exprimés. Les auteurs indiquent que les gènes qui ne partagent pas d'homologie et qui sont transcrits simultanément,

partagent des fonctions, et sont retrouvés proches à une fréquence deux fois supérieur à l'attendue. Dans cette étude, il est suggéré qu'il y a deux niveaux de régulation des gènes chez A. *thaliana* (Zhan et al., 2006). Un niveau de régulation local, où les éléments de régulation dupliqués en tandem ou bien qui partagent des éléments de régulation, contribuent à un fort niveau d'expression similaire pour un petit nombre de gènes voisins. Puis, à une échelle plus globale, de larges régions impliquant un état chromatinien ouvert ou fermé, et qui donc diffèrent entre les conditions expérimentales, ce qui peut expliquer le faible niveau de co-expression entre les grands groupes de gènes. Ce qui implique que les plantes vont pouvoir s'adapter aux changements d'environnement extrêmes (comme par exemple la sécheresse) (Zhan et al., 2006).

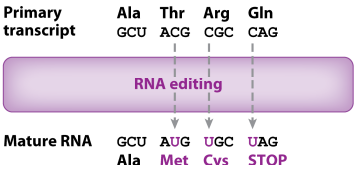
### 1.3.3 La duplication des gènes

La duplication des gènes est un mécanisme important de l'évolution et de l'adaptation des plantes à leur environnement. Plusieurs mécanismes peuvent être impliqués dans les duplications de gènes (Magadum, Banerjee, Murugan, Gangapur, & Ravikesavan, 2013) :

(i) les « cossing-overs » inégaux, qui conduisent à la formation de séquences répétées en tandem à partir de deux chromosomes homologues ou deux chromatides sœurs.
(ii) la rétrotransposition d'un transcrit.
(iii) la transposition duplicative : via la recombinaison homologue non allélique (NAHR) à partir de deux séquences homologues de deux chromosomes non homologues ou la réparation des cassures double brin via le mécanisme de « Synthetis Dependent Strand Annealing » (SDSA) (non homologous end joining, NHEJ). Cette voie a été suggérée suite à une étude sur le génome humain, montrant que dans certains cas il n'y avait pas d'implication d'une séquence répétée d'ADN ou bien de grandes parties de séquences homologues au niveau des points de cassures (Linardopoulou et al., 2005). La différence entre ces deux procédés réside dans l'utilisation de la séquence homologue durant la réparation de la cassure double brin.
(iv) la polyploïdisation.

Et les conséquences de la duplication d'un gène sont multiples : une copie d'un gène nouvellement dupliqué peut soit : conserver sa fonction, soit devenir non fonctionnelle (=pseudogénisation), soit être sélectionnée pour l'acquisition d'une nouvelle fonction (néo- ou sous-fonctionnalisation) (Magadum et al., 2013; Rensing, 2014). La néo-fonctionnalisation correspondant à l'acquisition d'une nouvelle fonction à partir de la duplication d'un gène. La sous-fonctionalisation correspondant à la conservation par un gène dupliqué d'une fonction ou bien d'un composant commun à la fonction originel (Comai, 2005).

A




B

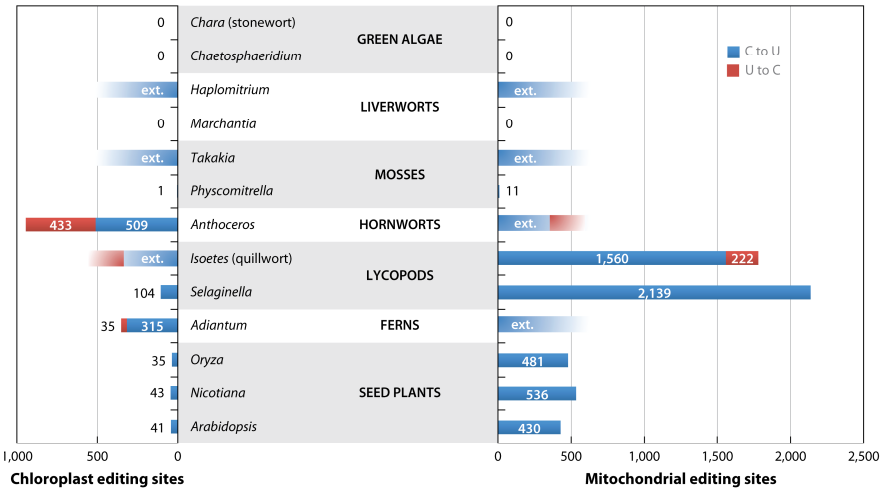



**Figure 10** : Edition de l'ARN entraine un changement de nucléotides.
A : La substitution de C en U entrain un changement d'acide aminé, et peut entrainer l'apparition d'un codon START ou STOP.
B : Chez les plantes, l'édition des ARN entraine le changement d'une base C en U dans les mitochondries et les plastes. Chez les algues vertes, l'édition des ARN n'a pas été observé. Chez les plantes hépatiques (ou « liverworts ») (Marchantiophyta), l'édition des ARN a été perdue. Le nombre de sites d'édition est donné pour chaque espèce. D'après (Takenaka et al. 2013).

Les gènes dupliqués en tandem représentent entre 10 et 20% du contenu en gènes dans les génomes du riz et A. *thaliana* (Rizzon, Ponger, & Gaut, 2006). Par ailleurs, une corrélation entre proportion de gènes dupliqués en tandem et taux de recombinaison méiotique a été décrite, suggérant un rôle majeur des mécanismes de recombinaison homologues dans la duplication de gènes (Rizzon et al., 2006). D'un point de vu fonctionnel, il a été montré que les gènes dupliqués en tandem les plus fréquemment maintenus dupliqués au cours de l'évolution étaient impliqués dans des fonctions extracellulaires et de réponses aux stress (Rizzon et al., 2006). Chez le soja (Glycine max), dont le génome a subit deux évènements de polyploïdisation différents (13 et 59 millions d'années), 75% des gènes ont été retrouvés présents en plus de une copie (Roulin et al., 2012). L'analyse de données RNA-Seq issues de 7 tissus, a permis de montrer que pour un sous set de 18 000 gènes, 50% des paralogues ont une expression différentielle et sont donc sous-fonctionnalisés. L'analyse de l'ontologie de ces gènes montre que seulement une petite proportion des gènes dupliqués issus de WGD ont été néo-fonctionnalisés (Roulin et al., 2012).

En ce qui concerne les gènes dupliqués en tandem par un mécanisme de duplication, chez le riz et A. *thaliana*, il a été montré qu'il y avait une corrélation positive entre la densité de gènes dupliqués en tandem et le taux de recombinaison (Rizzon et al., 2006).

# 2 La notion de gène et ses évolutions

## 2.1 Evolution de la notion de gène

Etymologiquement, le mot gène vient du grec *genesis* (« naissance ») ou *genos* (« origine »). C'est en 1909 que le terme de gène a été utilisé pour la première fois par le biologiste danois W. Johannsen, qui s'est basé sur le concept développé par G. Mendel en 1866. En effet avec ses travaux, G. Mendel avait montré que le phénotype était transmis de manière discrète à la descendance. En d'autres termes, que le phénotype est causé par le génotype. On commence à attribuer un gène à un locus dans les années 1910, avec les travaux de T.H Morgan sur la Drosophile avec la ségrégation de mutations, où il définit le modèle suivant : les gènes sont organisés de manière linéaire et la capacité de « cross-over » est proportionnelle à la distance qui sépare les gènes (Morgan, Sturtevant, Muller, & Bridges, 1915).

Du point de vue structural, le gène est défini comme une portion du génome composée d'introns et d'exons, produisant un ARN messager (noté ARNm) polyadénylé, pouvant coder une protéine, ou bien être transcrit en ARN non codant. Crick en 1958 propose l'analyse

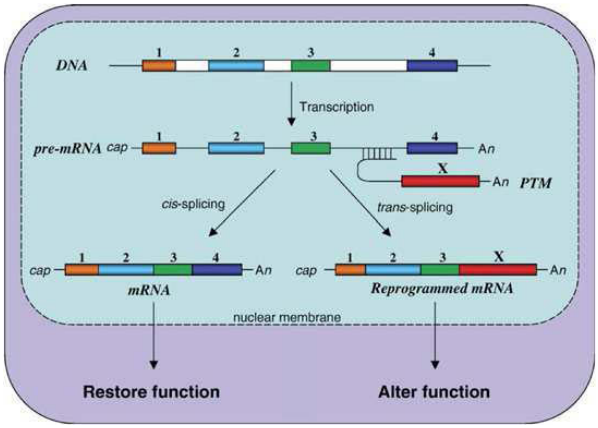



**Figure 11** : Représentation schématique de l'épissage en *cis* et *trans.*
Exemple d'un gène à 4 exons. L'ARN pré-messager contient les séquences introniques et exoniques. A gauche, l'épissage en cis génère un ARNm contenant les 4 exons du gène de départ. A droite, l'épissage en trans permet l'incorporation d'un exon X à la place de l'exon n°4. (Yang et Walsh 2005)

suivante : le flux de l'information de l'expression d'un gène provient d'un acide nucléique pour former une protéine. Cette analyse forme le « dogme central » qui décrit un système expliquant comment l'information stockée dans une séquence d'ADN est transférée du génome en une protéine fonctionnelle. Mais est-ce que tout se résume à la simple équation : un gène = une protéine ? La réponse à cette question est fournit en 1977 avec la mise en évidence de l'épissage alternatif par Berget et ses collaborateurs, qui décrit qu'un gène peut donner plusieurs transcrits (Berget, Moore, & Sharp, 1977). La découverte de l'épissage a complexifié le concept de gène, qui est définit alors comme pouvant donner naissance à un groupe de transcrits partageant un set d'exons dont un au moins est commun à tous les transcrits.

En 1986, une nouvelle découverte va encore remettre en question le dogme central précédemment décrit. Chez le trypanosome, l'insertion de 4 bases U dans le transcrit *cox2* présent dans les mitochondries permet de rétablir le cadre de lecture initialement absent du gène (Benne et al., 1986). L'information initialement portée par l'ARN devient donc différente de celle présente dans le génome. D'autres phénomènes d'insertions ont été décrits par la suite. Chez les kinétoplastidés (protistes), l'insertion de U peut représenter 55% des séquences d'ARNm produites (Stuart, 1991). Ces modifications de la séquence de l'ARNm qui ne sont pas retrouvées dans la séquence ADN correspondante sont appelées : édition des ARN ou « RNA editing ». C'est une modification post-transcriptionnelle de l'ARN qui change le contenu de l'information génétique. En effet, ce mécanisme entraine la substitution d'une base U (présente dans la séquence ARN), à la place d'une base C dans la séquence ADN correspondante, entraînant le changement de l'acide aminé de l'ARN mature (Figure 10A) (Covello & Gray, 1989; Takenaka, Zehrmann, Verbitskiy, Härtel, & Brennicke, 2013). Chez les angiospermes, ce phénomène a été mis en évidence pour la première fois en 1989 dans la mitochondrie avec des différentes de séquences entre l'ADN et l'ARN (Covello & Gray, 1989). Aujourd'hui, le phénomène d'édition des ARN n'a pas encore été mis en évidence dans les ARN cytoplasmiques (Takenaka et al., 2013) (Figure 10B).

Cependant, tous les gènes ne codent pas une protéine, comme par exemple les grands ARN non codant (sens ou anti-sens). En effet, si ces ARN peuvent être transcrits, ils ne codent pas de protéine. Cependant, des exceptions existent, il a été décrit certains loci peuvent être transcrits en grand ARN non codant ou bien coder de multiple petits peptides. Ces gènes sont généralement impliqués dans la régulation de l'expression de gènes codant des protéines (Kageyama, Kondo, & Hashimoto, 2011).

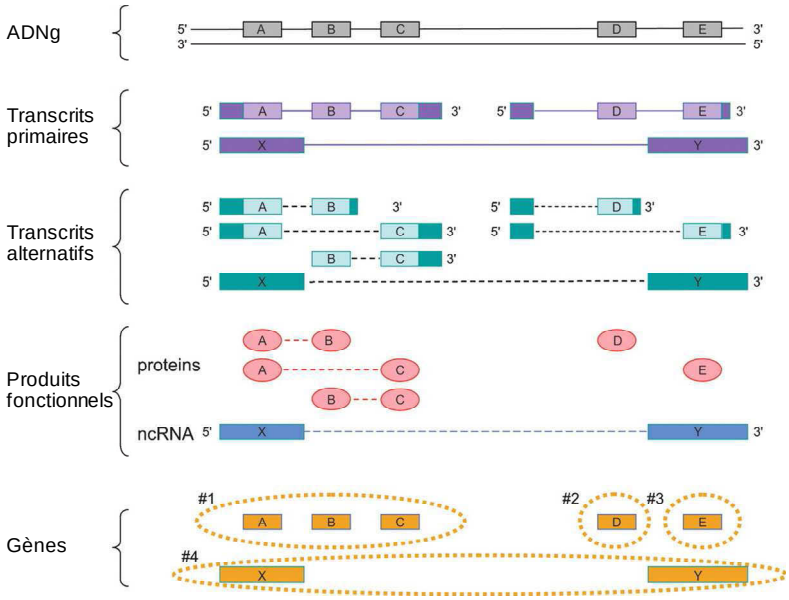



**Figure 12** : Illustration de la nouvelle définition d'un gène.
Dans cette région, 4 gènes sont représentés par des segments (A-Y) entourés d'une ligne en pointillés oranges. Les rectangles de couleurs pleines sont les régions non traduites, les rectangles de couleurs pastelles représentent les régions traduites. Les transcrits alternatifs sont représentés par des lignes vertes en pointillés. Deux transcrits primaires partagent des régions non traduites (X et Y). D'après (Gerstein et al. 2007)

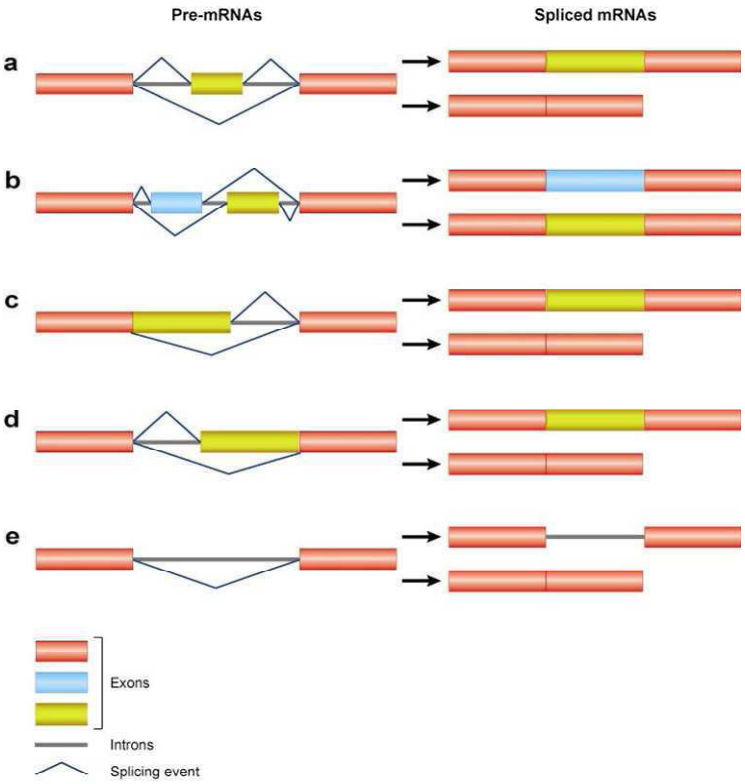



**Figure 13 :** Exemples de différents types d'épissage alternatif.
A gauche est représenté l'ANR pré messager et à droite des différents ARNm épissés. Les exons sont représentés par les rectangles colorés et les introns par des lignes horizontales. Les différentes formes sont :
(a) saut d'exon
(b) exclusion alternative d'un exon,
(c) épissage alternatif en 3'
(d) épissage alternatif en 5'
(e) rétention d'intron
D'après (Reddy 2007)

Un autre phénomène a aussi participé à la complexification du concept de gène : l'épissage en *trans* (« *trans*-splicing »). Il consiste au rattachement de molécules d'ARNm transcrites à partir de loci éloignés dans le génome (Blumenthal, 2005) (Figure 11).

Si l'on résume, la définition commune d'un gène portant sur l'habilité à déterminer un caractère particulier d'un organisme, ainsi que son héritabilité, a évolué. Le consortium « Sequence Ontology Consortium » définit aujourd'hui un gène comme « une région repérable de la séquence génomique correspondant à une unité héritée, transcrite, et associée à des régions régulatrices et/ou d'autres séquences fonctionnelles. ». Toutefois, cette définition est discutable pour plusieurs raisons. La première concerne les unités de régulation du gène (nécessaires pour la transcription d'un gène), plus particulièrement celles situées à longue distance. D'un point de vue de biologie moléculaire, les éléments de régulation peuvent être intégrés à la définition du gène sous la forme suivante : une séquence d'acide nucléique entière qui est nécessaire pour la synthèse d'un polypeptide fonctionnel ou d'un ARN (Lodish et al., 2000). Le second point réside dans le chevauchement des gènes. En effet, un gène peut être retrouvé inclue dans l'intron d'un autre gène ou bien un gène peut chevaucher un autre gène dans le même sens sans pour autant partager d'exon ou d'éléments régulateurs.

Ainsi, avec les problèmes posés par la précédente définition d'un gène et l'augmentation des connaissances acquises grâce au séquençage et l'annotation de nombreux génomes, une nouvelle définition est proposée, afin de refonder la définition du gène en prenant en compte cinq critères (Gerstein et al., 2007) : (i) la nouvelle définition doit être compatible avec l'ancienne, car un gène déjà définit doit pouvoir être retrouvé avec la nouvelle définition, (ii) elle doit être indépendante de l'organisme considéré, c'est-à-dire compatible avec les procaryotes et les eucaryotes, (iii) elle doit être le reflet d'une idée et non l'énumération de plusieurs mécanismes ou particularités, (iv) elle doit être simple à énoncer, et à mettre en œuvre pour répondre facilement à une question comme par exemple : « combien y'a t il de gènes dans le génome du blé ? », (v) enfin elle doit être compatible avec les autres nomenclatures biologiques déjà définies qui utilisent le concept de gène, comme par exemple le régulome : qui représente l'ensemble des interactions de régulation dans un organisme. Suivant ces considérations, (i) un gène est une séquence génomique qui codent pour des molécules fonctionnelles (ARN ou protéines), (ii) dans le cas où les produits fonctionnels partageraient des régions géniques, ce serait l'union de toutes les séquences chevauchantes codant pour le gène qui serait prise en compte, (iii) et cette union doit être cohérente mais ne signifie pas que tous les produits partagent une sous séquence

commune. Pour résumer: « un gène est l'union de séquences génomiques codant un ensemble cohérent de produits fonctionnels potentiellement chevauchants » (Figure 12).

Pour définir un gène, il est possible de considérer l'aspect génétique au sens propre du terme, où le gène représente une unité porteuse d'une information génétique transmise à la descendance et déterminant un caractère phénotypique. Le gène peut aussi être considéré du point de vue strictement moléculaire dans lequel le gène est un fragment d'ADN qui code un ARN pouvant ou non être traduit en protéine.

## 2.2 Epissage alternatif

### 2.2.1 Mécanismes

L'épissage alternatif est un processus qui consiste à produire deux ou plusieurs ARNm différents, appelés isoformes, à partir d'un même gène. Ce phénomène se produit par l'utilisation alternative des sites d'épissage en 5' et 3' dans le spliceosome (Roy, Haupt, & Griffiths, 2013). Ce processus correspond à une série d'évènements de conformation, interrompus par des épisodes chimiques de séparation des jonctions intron/exon, et du regroupement des exons entre eux. Ce mécanisme est contrôlé par des facteurs d'épissage de l'ARN pré-messager au moment de la formation du spliceosome. Parmi ces facteurs, on peut trouver une classe de protéines particulières : les protéines riches en acides-aminés sérine et arginine (SR), qui possèdent un ou deux domaines de fixation à l'ARN ainsi qu'un domaine d'interaction protéine/protéine. La sélection des facteurs d'épissage employés par chaque spliceosome dépend de la concentration de ces facteurs dans chaque type cellulaire et de la régulation des éléments présents pour chaque ARN pré-messager. Suite à un évènement d'épissage alternatif, deux solutions sont possibles pour le produit qui en résulte : (i) la régulation du niveau de transcrits, par la production d'une isoforme instable, qui pourra être dégradée par la voie « non-sens mediated decay », un microARN ou une autre voie de dégradation, (ii) la production d'un transcrit fonctionnel, qui conduira à la production d'une isoforme de protéine qui diffère au niveau de la localisation subcellulaire, de la stabilité ou bien de la fonction, de la forme constitutive du gène (Barbazuk, Fu, & McGinnis, 2008).

Sept types majeurs d'épissage alternatif ont été décrits : saut d'exon, rétention d'intron, épissage en 5', épissage en 3', promoteur et premier exon alternatif, exclusion alternative mutuelle d'exon, site de poly-adénylation et exon terminal alternatif (Reddy, 2007; Syed, Kalyna, Marquez, Barta, & Brown, 2012) (Figure 13). Les phénomènes d'épissage alternatif sont plus abondants chez les grands eucaryotes que chez les petits, et le pourcentage de gènes ainsi que d'exons qui subissent un épissage alternatif est plus fort chez les vertébrés

que chez les invertébrés (Frankish, Mudge, Thomas, & Harrow, 2012). Chez l'homme, des études ont montré que plus de 80% des gènes subissent des évènements d'épissages alternatifs (Barbazuk et al., 2008). Chez les plantes, le pourcentage est plus faible et représente environ 60% des gènes (Barbazuk et al., 2008). Deux études ont montré qu'il n'y a pas de différence dans les séquences consensus des sites d'épissage 5' et 3' entre les animaux et les plantes, cependant, des différences ont été observées au niveau des introns en ce qui concerne la séquence des points de branchement de la boule d'épissage, la taille, ainsi que la composition en nucléotides (Reddy, 2007; Iwata & Gotoh, 2011). Différences ont aussi été décrites au niveau des évènements initiaux de l'étape d'épissage impliquent une reconnaissance d'un site d'épissage unique pour les plantes (Reddy, 2007). Cependant, chez les plantes, la présence de séquences riches en uridine aux alentours du site d'épissage en 3' a été mise en évidence comme un facteur important pour l'efficacité de l'épissage alternatif.

Les sites d'épissage sont de courtes séquences nucléotidiques entourant la frontière intron-exon, et sont cruciaux pour l'épissage de l'ARN pré-messager. Pour la reconnaissance des sites d'épissage, les courtes séquences consensus pour les sites donneur et accepteur sont nécessaires mais ne sont pas suffisantes. Des études sur des génomes d'animaux ont montré que d'autres séquences, exoniques et introniques, régulatrices en *cis* ou *trans*, influencent la sélection des sites d'épissage constitutifs et alternatifs (Reddy, 2007). Deux modèles sont ainsi définis pour expliquer la reconnaissance des sites d'épissage : la préférence de l'exon et la préférence de l'intron. Le modèle préférence de l'exon implique que la machinerie assemble les exons par la reconnaissance en premier des sites d'épissage aux alentours des exons. A l'inverse c'est la séquence régulatrice intronique du site d'épissage qui est reconnue par la machinerie pour le modèle intron (Berget, 1995). Concernant le génome de l'homme ainsi que ceux d'autres vertébrés où les gènes sont composés de petits exons séparés de grands introns, c'est le modèle préférence de l'exon qui est favorisé (Reddy, 2007). Chez les plantes, les exons ont un fort pourcentage de bases GC, et aucune séquence spécifique aux exons impliquée dans la reconnaissance ou la régulation au niveau du site d'épissage n'a encore été identifiée. De plus, les gènes des plantes sont connus pour avoir de courtes séquences introniques (~150 pb), à la différences des génomes animaux, où la taille des introns est en moyenne de 5 kb (Sakharkar, Chow, & Kangueane, 2004; Wendel et al., 2002). Le modèle préférence de l'intron est le plus adapté pour les génomes des organismes ayant des gènes avec de petits introns (comme les génomes des plantes). Cette hypothèse est soutenue par le fait que la forme de rétention d'intron est prédominante dans les génomes de plantes (Syed et al., 2012).

### 2.2.2 Implications biologiques

Chez Arabidopsis et le riz, les événements d'épissage alternatifs ont été analysés à partir d'alignement de séquences EST et d'ADNc. Les résultats montrent que, 22% et 21% des gènes subissent au moins un évènement d'épissage alternatif respectivement chez ces deux espèces (B.-B. Wang & Brendel, 2006). Alors que la forme de saut d'exon est la forme majoritairement retrouvée dans le génome humain avec 58% des événements d'épissage, chez Arabidopsis et le riz c'est la forme de rétention d'intron qui est majoritaire avec 56% et 54% des évènements respectivement. La forme « saut d'exon » ne représente que 10% des évènements en moyenne pour ces deux génomes. Dans cette étude, les auteurs ont aussi analysé la conservation des évènements d'épissage entre Arabidopsis et le riz. Les résultats montrent que 40% des gènes homologues entre les deux espèces sont épissés. Toutefois, l'utilisation de séquences EST et ADNc présentes dans les bases de données limite la capacité à identifier tous les événements, notamment les isoformes rarement et/ou peu produites par la cellule. A partir de donnée issues de séquençage d'ARNm, les évènements d'épissage alternatif ont été aussi investigués chez d'autres génomes de plantes, comme le concombre (S. Guo et al., 2010), la vigne (Zenoni et al., 2010), et Brachypodium (Walters, Lum, Sablok, & Min, 2013). Toutes ces études ont montré que en moyenne plus de 60% des gènes contenant des introns pouvaient être épissés.

Afin de démontrer l'implication biologique des transcrits issus d'évènement d'épissage alternatif, six observations ont été rapportées (Reddy, 2007) :

1/ Si les isoformes sont le fruit d'erreurs du processus d'épissage, alors tous les gènes possédants au moins un intron pourraient entrainer la production d'un transcrit alternatif. Cependant, l'épissage alternatif est prédominant dans certaines familles de gènes alors que dans d'autres familles, où la taille et le nombre d'intron est similaire, il n'y a pas de production de transcrits alternatifs.

2/ Dans beaucoup de cas, les évènements d'épissage se produisent dans des familles de gènes qui codent des protéines multi-domaines. Les isoformes qui en résultent diffères dans l'organisation des domaines et sont, par conséquent, susceptibles d'avoir des fonctions différentes.

3/ Chez les plantes, il a été montré que l'épissage alternatif est régulé de manière tissu-spécifique, à des stades de développement précis et en réponse à des conditions de stress (Shi, Xiong, Stevenson, Lu, & Zhu, 2002; Yoshimura, Yabuta, Ishikawa, & Shigeoka, 2002).

4/ La position des introns alternatifs est conservée entre les gènes de plantes distantes au niveau évolutif (Arabidopsis et riz par exemple) (Iida & Go, 2006; Masayuki Isshiki, Tsumoto, & Shimamoto, 2006; B.-B. Wang & Brendel, 2006).

5/ Les isoformes avec des introns sont recrutées pour la traduction.

6/ Beaucoup d'ARNm issus du processus rétention d'intron ne sont pas éliminées par les mécanismes de contrôle de l'ARN dans le cytoplasme.

Un exemple type de caractère connu pour être contrôlé par l'épissage alternatif est la production d'amylose chez le riz, les riz cultivés ayant un taux d'amylose plus faible que celui des espèces sauvages. Cette différence s'explique par une mutation dans un site d'épissage en 5' du gène *Waxy*, affectant l'épissage alternatif de l'ARN pré-messager (M Isshiki et al., 1998).

Alors que la rétention d'intron est la forme d'épissage alternatif la plus fréquente chez les eucaryotes inférieurs, la prévalence de la forme « saut d'exon » augmente graduellement avec la complexité des organismes eucaryotes (Keren, Lev-Maor, & Ast, 2010). En revanche, chez les plantes, la forme rétention d'intron reste majoritaire, à l'inverse de la forme « saut d'exon » qui reste rare (Syed et al., 2012). Ces informations suggèrent que, contrairement aux vertébrés, chez les plantes l'épissage alternatif n'a pas été un mécanisme aussi majeur dans l'évolution du potentiel codant. L'importance des duplications de gènes, notamment via la polyploïdie, pourrait être une explication de cette différence. Afin d'étudier la « viabilité » des transcrits issus d'épissage alternatif, la proportion des événements au sein des régions codantes a été observée (Filichkin et al., 2010). Chez Arabidopsis, des données obtenues par RNA-Seq montrent que ~78% des évènements d'épissages impliquent la formation d'un codon stop prématuré, ce qui fait des ces isoformes une cible privilégiée de la voie de dégradation « non-sense mediated mRNA decay » (NMD) (Filichkin et al., 2010; Marquez, Brown, Simpson, Barta, & Kalyna, 2012; Soergel, Lareau, & Brenner, 2006).

## 2.3 Les pseudogènes

### 2.3.1 Définition et classification des pseudogènes

Les gènes ne sont pas les seules entités transcrites dans un organisme. Les pseudogènes (notés Ψ) ont été mis en évidence en 1977 chez le Xénope avec la découverte d'une version tronquée du gène codant pour l'ARNr 5S et présentant une forte homologie avec le gène actif dans l'organisme (Jacq, Miller, & Brownlee, 1977). Par la suite, ils ont été identifiés aussi bien chez les procaryotes que chez les eucaryotes. Les pseudogènes sont

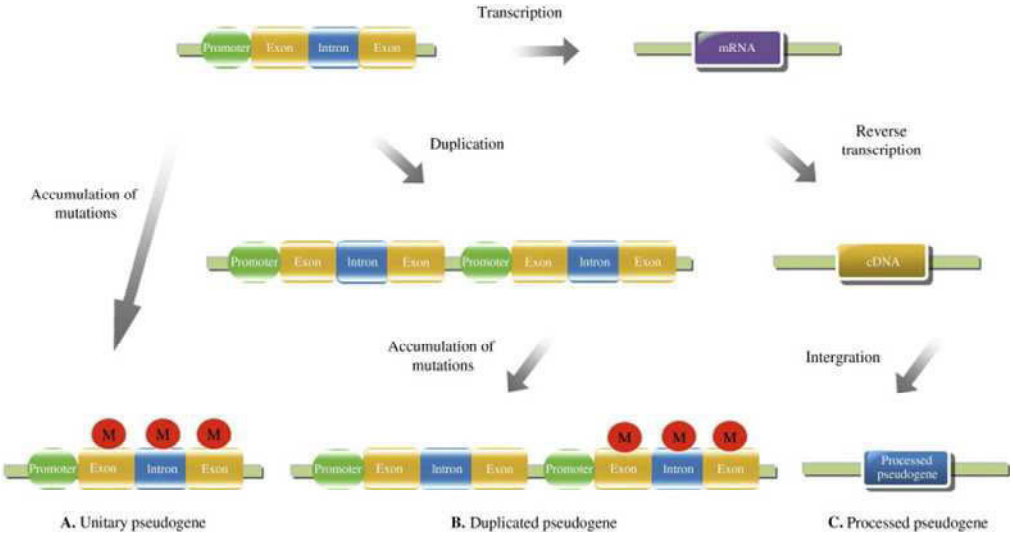



**Figure 14** : Règles de classification des pseudogènes.
A : pseudogène unitaire, B : pseudogène dupliqué, C : rétro-pseudogène. La lettre M indique les mutations. D'après (W. Li, Yang, et Wang 2013)

traditionnellement définis comme des séquences génomiques qui ressemblent aux séquences de gènes codant une protéine, mais qui ne sont pas fonctionnelles (Zheng & Gerstein, 2007). Cela pour plusieurs raisons : l'absence de promoteurs, la présence de mutation non-sens, un décalage du cadre de lecture, la perte de site d'épissage ou toutes autres mutations délétères (Balakirev & Ayala, 2003; W. Li, Yang, & Wang, 2013; Pei et al., 2012; Zheng & Gerstein, 2007). La frontière entre les gènes et pseudogènes est, en réalité, difficile à définir, et des études ont montré que certains pseudogènes (d'un point de vue structural) ont une fonction, souvent impliquée dans la régulation de l'expression du gène parent (Zheng & Gerstein, 2007). Ainsi, pour prendre en considération cette complexité, les pseudogènes sont définis comme une séquence génomique qui provient d'un gène parent fonctionnel mais qui n'a pas la même fonction que le gène parent (Zheng & Gerstein, 2007).

Les pseudogènes peuvent être issus de la dégradation d'un gène codant une protéine, présent en copie unique ou dupliqué au sein du génome, ou de la rétro-transposition d'un ARNm (W. Li et al., 2013 ; Balakirev & Ayala, 2003). Ainsi, on distingue (Figure 14) :

- les pseudogènes unitaires qui résultent de l'accumulation de mutations dans un gène en copie unique.
- les pseudogènes dupliqués qui dérivent de duplications de gènes parents fonctionnels, mais dont une des copies est inactive.
- les rétro-pseudogènes qui proviennent de la réverse transcription d'un ARNm (non apparenté à un élément transposable) et de l'intégration de l'ADNc dans le génome.

A la différence des rétro-pseudogènes, les pseudogènes dupliqués peuvent avoir maintenu les séquences régulatrices en amont, issues de leurs parents.

En ce qui concerne l'origine des pseudogènes, une différence a été observée sur la fréquence de production des pseudogènes pour différents gènes (Pei et al., 2012). Les gènes de ménage, qui sont fortement exprimés dans les cellules germinales et qui participent au métabolisme basal, ont tendance à générer un grand nombre de pseudogènes dans les génomes. Une explication simple serait que ces gènes fortement exprimés ont davantage de chance d'être rétro-transposés. De plus, le taux de GC affecte le taux de mutation (Bustamante, Nielsen, & Hartl, 2002). Ainsi, l'apparition de pseudogènes serait fonction du taux de GC du gène. Un autre facteur à prendre en compte est la taille des gènes. Il a été montré que la taille des gènes impacte sur le type de pseudogène produit, les gènes de grande taille sont associés à des pseudogènes unitaires et dupliqués, alors que les gènes de petite taille sont associés aux rétro-pseudogènes. Dans le génome humain, les

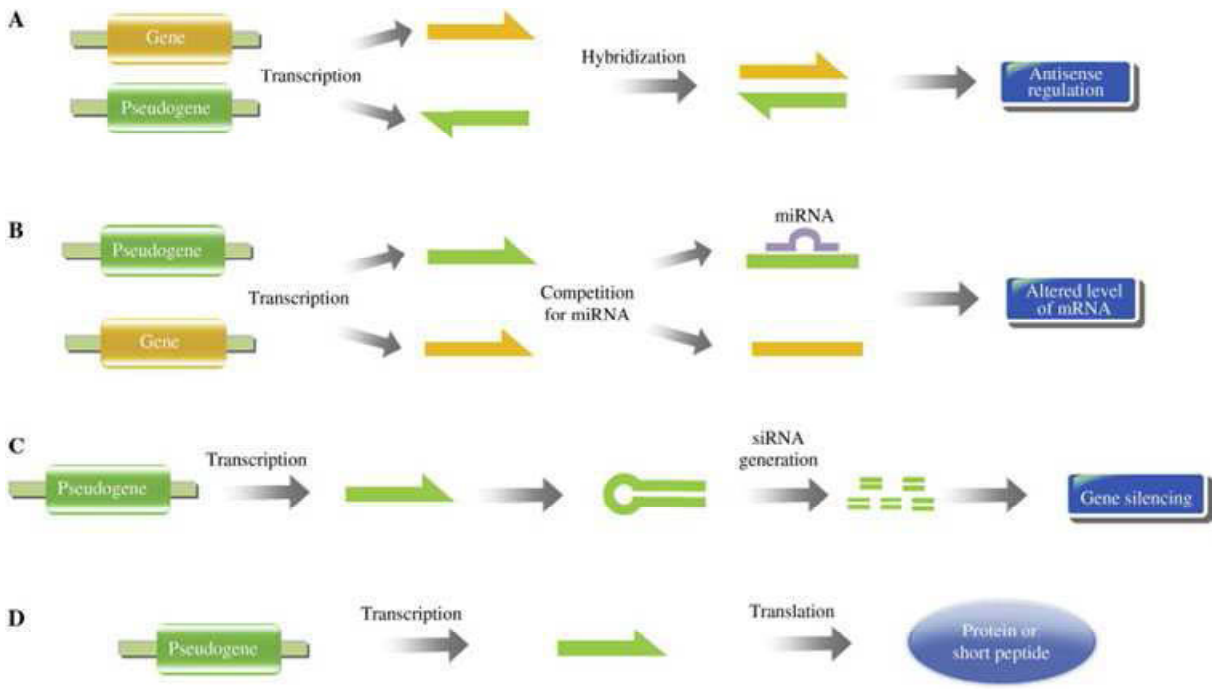



**Figure 15** : Différents modèles de fonction des pseudogènes.
A : Un pseudogène est reverse transcrit, et va réguler l'ARNm parent via un mécanisme de régulation anti-sens.
B : Le pseudogène transcrit va servir d'appât au miRNA et va donc empêcher le ciblage du gène par le miRNA.
C : Après la transcription, le pseudogène avec des séquence répétées inversées va former une structure secondaire en épingle à cheveu et produire des siRNA endogènes.
D : certains pseudogènes vont acquérir la capacité de coder des petits peptides ou bien des protéines, souvent avec de nouvelles fonctions.

gènes parents des rétro-pseudogènes sont largement exprimés, fortement conservés, de petite taille, et avec un pourcentage en GC faible (Gonçalves, Duret, & Mouchiroud, 2000).

### 2.3.2 Identification des pseudogènes

Si on se réfère à la définition d'un pseudogène, sa séquence doit avoir un fort taux d'identité avec celle du gène parent, sans nécessairement avoir une structure fonctionnelle (intron/exon/cadre de lecture). L'identification des pseudogènes est une étape importante dans l'annotation d'un génome, et malheureusement un grand nombre de pseudogènes peuvent être confondus avec leur parent fonctionnel. Pour les pseudogènes unitaires, l'identification est basée sur la comparaison de séquence avec un orthologue présent chez d'autres génomes apparentés. Avec l'augmentation des données d'expression de gène via le séquençage, l'identification des pseudogènes ne cesse de croitre (Pei et al., 2012). Toutefois, une des limites est de pouvoir distinguer l'expression du pseudogène et de celle de la copie parente. C'est notamment le cas des pseudogènes dupliqués récemment et n'ayant que très peu divergé du gène parent.

Chez Arabidopsis et le riz, 3 719 et 7 902 pseudogènes ont été mis en évidence respectivement, représentant 11% et 22% des gènes/pseudogènes annotés (Zou et al., 2009). L'analyse de l'expression de ces pseudogènes a montré que le niveau est en plus faible proportion que les gènes fonctionnels et qu'ils résultent majoritairement de duplications récentes.

### 2.3.3 Pseudogènes: transcription et fonction

Les pseudogènes peuvent réguler l'expression du gène parent (Figure 15). Pour cela, quatre modes d'action ont été mis en évidence. Tout d'abord, certains pseudogènes peuvent être des transcrits anti-sens comme par exemple : pseudo-NOS (nitric oxyde synthase) décrit en 1999 chez l'escargot (Korneev, Park, & O'Shea, 1999). Pseudo-NOS contient une région anti-sens de l'ARNm issu du parent *nNOS* et peut donc éteindre l'expression de *nNOS* via la formation d'ARN double-brin (Zheng & Gerstein, 2007). Les pseudogènes peuvent aussi être ciblés par des miRNA et réguler l'expression du gène parent par compétition de la fixation des miRNA entre gène et pseudogène. Le troisième mode d'action est la production de siRNA par l'intermédiaire d'un ARN double-brin à partir d'un pseudogène réverse transcrit, entrainant la formation d'une structure secondaire en épingle à cheveux qui va générer des siRNA. Ces derniers seront capables de réguler l'expression du gène parent par ARN interférence. Ce phénomène a été montré chez les plantes, notamment chez le riz (X. Guo, Zhang, Gerstein, & Zheng, 2009) et A. *thaliana* (Kasschau et al., 2007). Enfin, certains

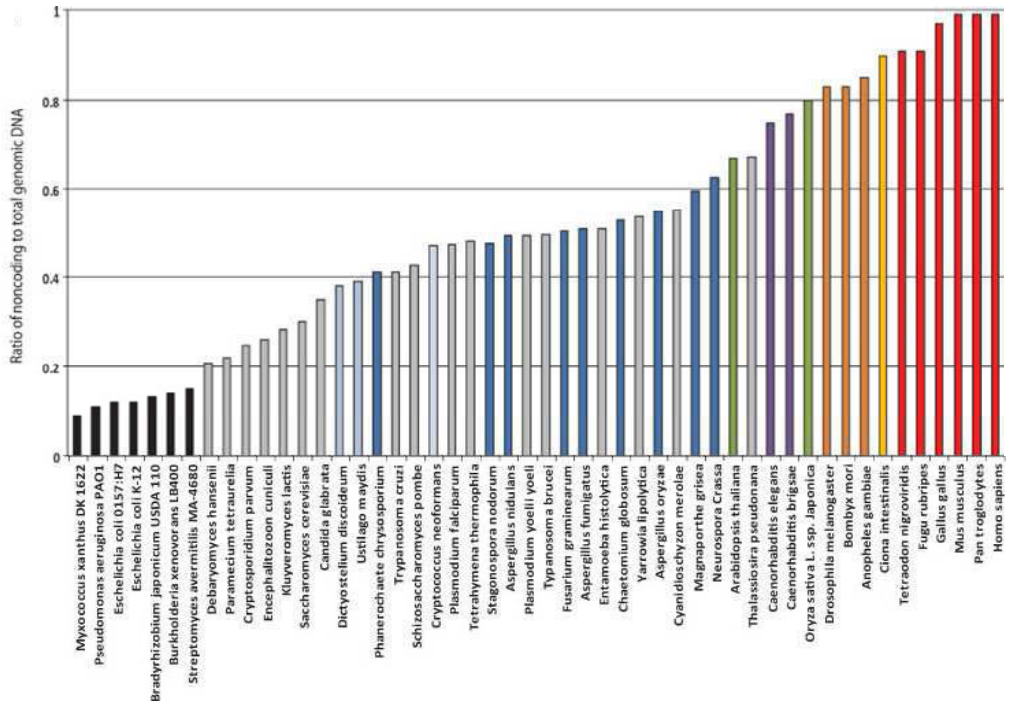



**Figure 16** : Fraction des séquences d'ADN qui ne codent pas pour une protéine, par génome haploïde dans différentes espèces.
En ordonnée est représenté le ratio du nombre total de bases composant les loci ne codant par pour des protéines, par rapport au nombre total de bases d'ADNg par génome pour chaque espèce (i.e. le pourcentage d'ADN non codant). En noir sont représentés les 4 plus grands génomes procaryotes ainsi que deux génomes bactériens bien caractérisés. Les organismes unicellulaires sont représentés en gris. En bleu clair, ce sont les organismes unicellulaire ou pluricellulaire en fonction du cycle de vie. En bleu foncé, on retrouve les organismes pluricellulaires de base. Les plantes sont représentées en vert. En violet ce sont les nématodes, les arthropodes sont en oranges, les chordates en jaunes et les vertébrés en rouges. Le nom de chaque espèce est représenté en bas de chaque barre. D'après (Taft, Pheasant, et Mattick 2007).

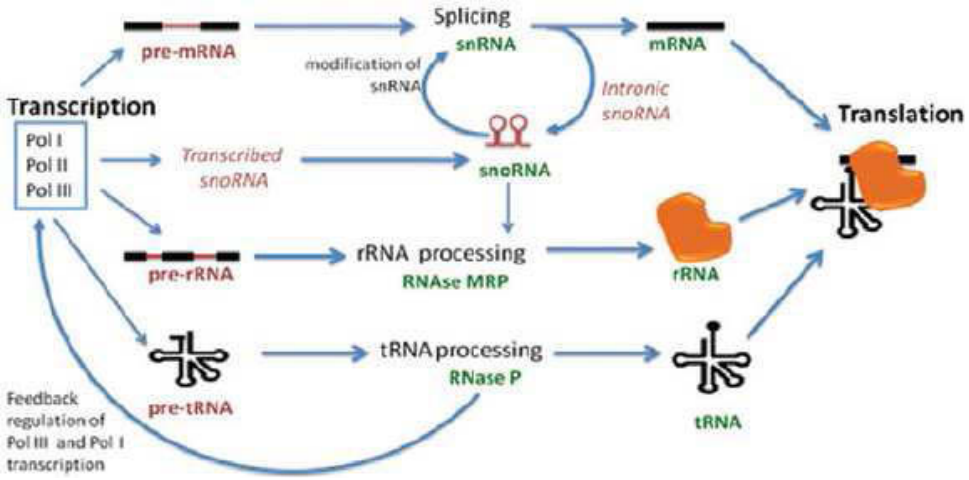



**Figure 17** : Schéma de la formation des ARNnc et de leurs intéractions (Eucaryotes).
Les ARNnc sont impliqués dans les mécanismes de maturation et de traduction des ARNm. Les snoRNA (petits ARN nucléolaires) sont impliqués dans la maturation des snRNA (petits ARN nucléaires) ainsi que dans l'épissage alternatif des ARNm. Les ARNr entrent en action lors de la traduction des ARNm, ainsi que les ARNt. D'après (Collins 2011).

pseudogènes peuvent être transcrits et traduits en une protéine tronquée. Par exemple chez l'homme, le pseudogène *Cx43* (Connexin43) code une protéine de 43 kDa dans des cellules tumorales, qui est capable d'inhiber la croissance cellulaire (Kandouz, Bier, Carystinos, Alaoui-Jamali, & Batist, 2004).

## 2.4 Les ARN non codant

Le terme d'ARN non codant (noté ARNnc) est employé pour définir les ARN qui ne codent pas une protéine. Cependant, ils peuvent posséder une activité catalytique qui leur est propre. Certains sont impliqués dans des complexes nucléoprotéiques (les ribosomes notamment), comme par exemple les miRNA qui font partie intégrante du complexe RISC (« RNA-induced silencing complex ») (Gurtan & Sharp, 2013). Dans les génomes des mammifères, des analyses transcriptomiques ont montré que deux tiers de l'ADN génomique était transcrit, ce qui contraste avec le pourcentage estimé à moins de 2% des gènes codants des protéines (Djebali et al., 2012). Par ailleurs, le degré de complexité entre les espèces a une plus forte corrélation avec la proportion des transcrits non codants, qu'avec les transcrits codants des protéines, même en prenant en compte la diversité des protéines issues d'épissage alternatif et modifiées par les mécanismes de régulation post-transcriptionnelle (Taft, Pheasant, & Mattick, 2007) (Figure 16). Cela suggère que les mécanismes de régulation impliquant des ARN ont un rôle important dans l'évolution de la complexité du développement chez les eucaryotes.

La proportion des ARNnc excède celle des gènes codants des protéines dans les génomes eucaryotes. On distingue les ARNnc dit constitutifs, c'est à dire ceux impliqués dans des mécanismes conservés comme la traduction (ARNr, ARNt), l'épissage ou la maturation des ARN (snRNA, snoRNA), et les ARNnc qui sont impliqués dans des phénomènes de régulation, généralement moins conservés entre les espèces (Figure 17).

Ces ARNnc sont moins bien caractérisés car plus difficiles à mettre en évidence. Cependant, le développement récent de protocoles de séquençage massifs d'ADNc a accru considérablement notre capacité à les identifier dans les génomes. Une grande proportion des ARNnc transcrits ont une taille supérieure à 200 pb. Ces ARN sont souvent polyadénylés mais sont dépourvus d'ORF (« open reading frame », cadre de lecture ouvert) et sont regroupés sous le terme de « long non coding RNA » ou grands ARN non codants (notés lncRNA) (Fatica & Bozzoni, 2014). Les gènes codant des lncRNA partagent des caractéristiques semblables aux gènes codants des protéines : profils de modification d'histone similaires, signaux d'épissage et tailles des introns et des exons similaires (Ulitsky & Bartel, 2013). Ils ont donc une composition proche de celle des ARNm mais ne servent

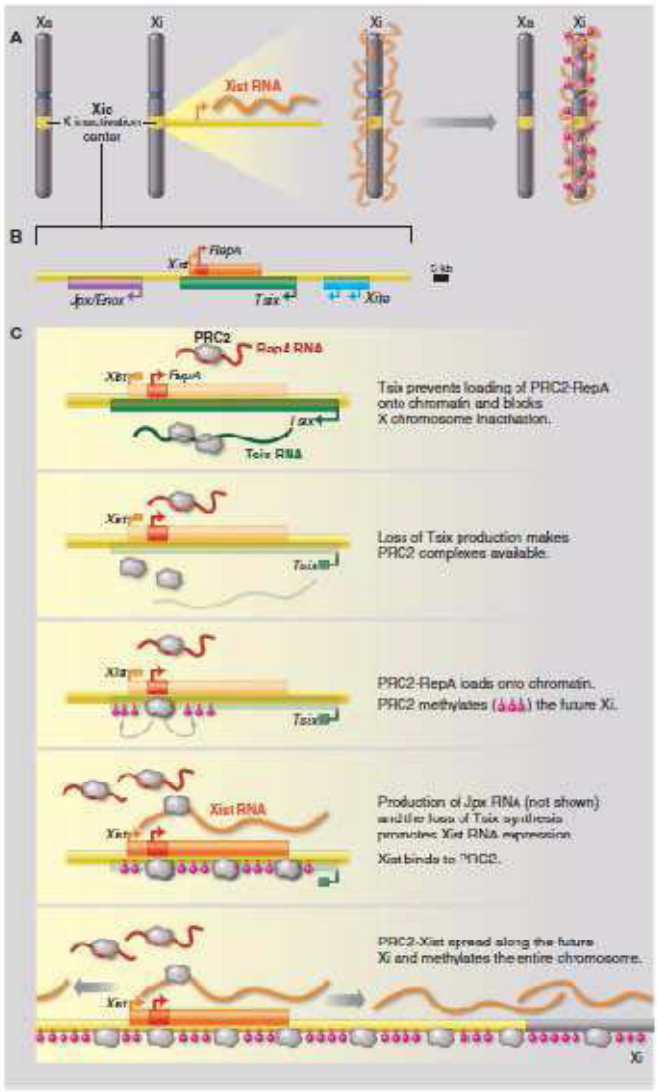



**Figure 18** : Inactivation du chromosome X.

(A) Le lncRNA Xist est trranscrit à partir du locus *Xic* situé sur le chrososome X inactif (Xi). L'ARN Xist recouvre entièrement le chromosome et entraine l'extinction de l'expression des gènes portés par le chromosome via des modifications épigénétiques des histones et de l'ADN.
(B) Le locus Xist et son lncRNA.
(C) Intéractions du lncRNA et des protéines lors de l'initiation de l'inactivation du chromosome X.
D'après (J. T. Lee 2012).

**Tableau 2** : Principales caractéristiques et fonctions des ARNnc. D'après (Dogini et al. 2014)

| | Classe | Taille | Fonctions |
|---|---|---|---|
| Long ncRNA | rRNA | ~1,9 kb | Essentiel pour la synthèse des protéines. |
| | XIST RNA | ~17 kb | Inactivation du chromosome X. |
| | Autres lncRNA | > 200 nt | Impliqué dans des modifications épigénétiques, processus post-transcriptionnels, modulation de la structure de la chromatine, etc ... |
| Small ncRNA | miRNAs | 18–21 nt | Régulation des gènes. |
| | siRNA | ~21 nt | Régulation des gènes, défense contre les virus et activité de transposon. |
| | rasiRNA | 24–27 nt | Orientation de l'hétérochromatine dans la formation du centromère. |
| | snoRNA | 60–300 nt | Méthylation et pseudo uridylation d'autres ARN |
| | snRNA | 100–300 nt | Impliqué dans le complexe du spliceosome. |
| | piRNA | 26–30 nt | Régulation de l'activité des transposons et de l'état chromatinien. |

pas de base à la synthèse de protéines (Tableau 2). Cependant, à la différence des ARNm, les lncRNA présentent un biais de composition en nombre d'exons (en faveur de deux exons) (Harrow et al., 2012). Chez l'homme, une analyse de l'expression dans 6 types cellulaires différents a montré que les lncRNA sont généralement moins exprimés que les gènes codant des protéines et sont davantage exprimés spécifiquement (Derrien et al., 2012). Cette spécificité est corrélée à la présence accrue d'éléments transposables au voisinage du promoteur des lncRNA (Kelley et Rinn 2012).

Quoiqu'encore peu décrit dans la littérature, un rôle biologique majeur a parfois été démontré pour certains ARNnc comme notamment l'inactivation d'un des chromosomes X chez les femmes, processus nécessaire au maintien de l'équilibre du niveau d'expression des gènes portés par ce chromosome. Le locus responsable de l'initiation de cette inactivation est transcrit en deux ARNnc appelés Xist (« X-inactive specific transcript ») et Tsix qui contrôlent la répression de l'expression de l'ensemble des gènes d'un des deux chromosomes X (K. C. Wang & Chang, 2011) (Figure 18).

Chez les plantes, plusieurs lncRNA ont déjà été décrits, majoritairement chez *A. thaliana*, comme par exemple COOLAIR qui est impliqué dans la vernalisation (Heo, Lee, & Sung, 2013). Au total, chez *A. thaliana* 2 708 lincRNA ont été identifiés (J. Liu et al., 2012). Plus récemment chez le maïs, un set robuste de lncRNA a été généré en se basant sur des données RNA-Seq de 30 tissus, ainsi que des EST (L. Li et al., 2014). Parmi les 20 163 lncRNA mis en évidence, 18 459 ont été identifiés comme des précurseurs de petits ARN et les 1 704 restants ont été définis comme un set lncRNA. En se basant sur plus d'un milliard de lectures RNA-Seq issus de 13 tissus de maïs, l'analyse des données a montré que 50% de lncRNA précédemment identifiés sont exprimés spécifiquement dans un tissu, alors que seulement 10% sont exprimés dans au moins cinq tissus (L. Li et al., 2014). Ces valeurs contrastent avec celles des gènes codant des protéines pour lesquels 8% sont exprimés dans un tissu uniquement et 74% dans au moins cinq tissus. En terme de niveau d'expression, 80% des lncRNA ont un FPKM inférieur à 5 dans chaque tissu testé, suggérant un niveau d'expression globalement faible des lncRNA (L. Li et al., 2014). C'est une des raisons qui explique qu'ils n'ont été mis en évidence que récemment, conjointement à l'augmentation des débits de séquençage.

## 3 Les outils d'analyse des gènes

Les avancées technologiques et méthodologiques, ainsi que les outils statistiques permettent à l'heure actuelle d'étudier les génomes à une échelle de précision jamais

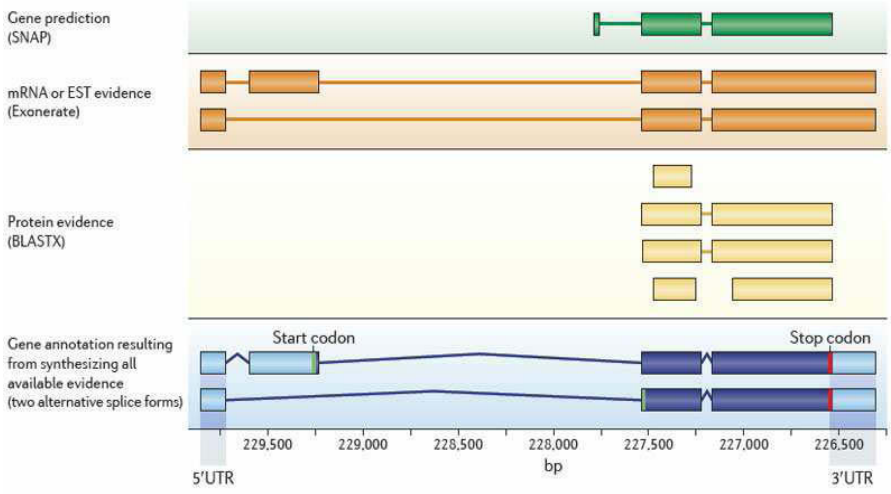



**Figure 19** : Différence entre prédiction et annotation d'un gène.
Le schéma représente l'annotation d'un gène ainsi que les évidences biologiques. Entre parenthèse ce sont les noms communément employés pour les programmes. En bleu est représentée l'annotation du gène avec les régions 5' et 3' UTR suggérées par les évidences biologiques (orange). La prédiction du gène qui est réalisée par SNAP (en vert) est incorrecte car il manque les exons en 5' ainsi que le site d'initiation de la traduction, et comme beaucoup de prédicteurs de gènes, il n'inclut pas les régions UTR. D'après (Yandell et Ence 2012).

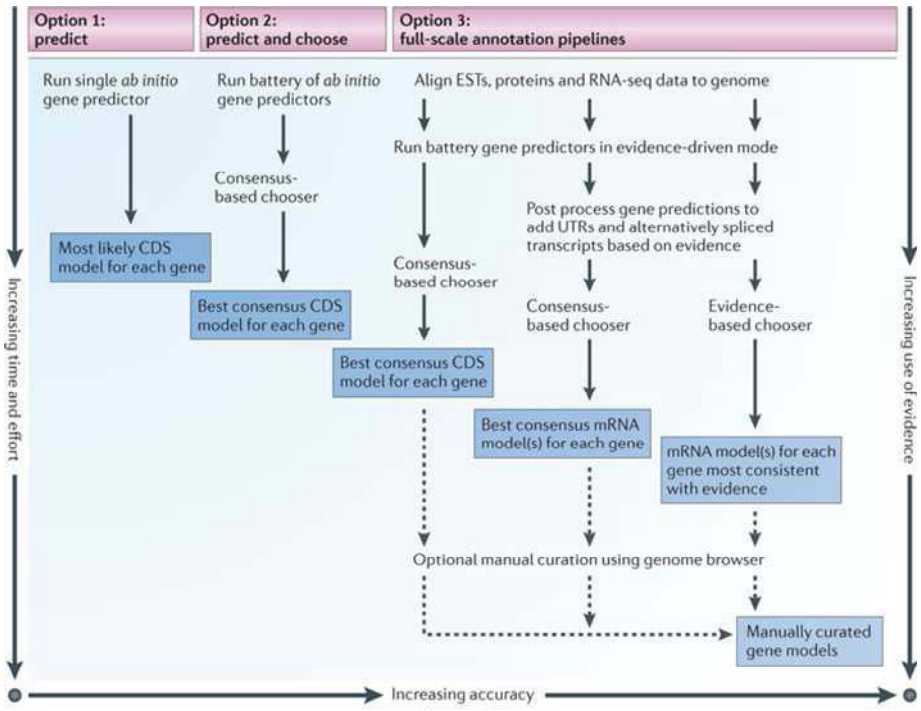



**Figure 20** : Schéma représentant trois approches basiques de l'annotation d'un génome.
Les trois approches différent en fonction du temps, de l'effort et de la qualité de l'annotation voulue, en regard d'une approche *ab initio*. Le produit final de chaque approche est indiqué dans les rectangles bleus.

atteinte jusqu'ici. Cela grâce aux outils d'annotation des gènes ainsi que de transcriptomique, qui permettent d'analyser les gènes tant au niveau de leur structure (taille, nombre d'exon, etc), que de leur fonction.

## 3.1 Annotation de l'espace génique

L'annotation structurale des gènes dans la séquence génomique consiste à identifier la présence de gènes et à prédire leur structure codante (pour les gènes codant des protéines). Une précision peut être apportée quant à la distinction à faire pour les termes « prédiction » et « annotation » de gènes (Yandell & Ence, 2012) (Figure 19). Les outils de prédiction de gènes recherchent la séquence codante du gène (CDS) et ne recherchent pas les régions UTR régulatrices ou les variants d'épissage. La prédiction de gènes est donc un terme trompeur, et le terme suivant pourrait être utilisé : « prédiction canonique des CDS ». Au contraire, l'annotation des gènes inclus les régions UTRs, ainsi que les isoformes issues de l'épissage alternatif.

Les pipelines d'annotation des génomes ont comme objectif de combiner automatiquement les informations issues des prédictions *ab initio* et des recherches de similarité, avec des banques de séquences de transcrits et de protéines, pour modéliser la structure du gène (Figure 20). De nombreux pipelines sont déjà disponibles pour l'annotation des génomes de plantes, comme par exemple : Megante (Numa & Itoh, 2014), RiceGAAs (Sakata et al., 2002), DAWG-PAWs (« DAWGPAWS - http://dawgpaws.sourceforge.net/ », s. d.). Pour le génome du blé, le pipeline TriAnnot a été développé dans l'Unité (Leroy et al., 2012). Cet outil prédit les positions des séquences codantes (CDS) et leurs fonctions potentielles, détermine l'aspect fonctionnel ou non (pseudogène) des gènes prédits et attribue un seuil de confiance quant à leur structure. Les données de séquençage de transcriptome sont aussi couramment utilisées aujourd'hui pour améliorer les prédictions de structure des gènes (Schatz, Witkowski, & McCombie, 2012).

## 3.2 Outils d'analyse de l'expression des gènes

De nombreuses méthodes ont été développées pour étudier les profils d'expression des gènes dans un tissu ou une cellule. Par exemple le Northern blot, la PCR après transcription inverse (RT-PCR), le séquençage d'EST (« expressed sequence tags »), le SAGE (« serial analysis of gene expression ») et les puces à ADN (plateforme Fluidigm (Frederickson, 2002), Affymetrix (« Whole Transcriptome | Expression Analysis | Affymetrix » 2014), Nimblegen (http://www.nimblegen.com/support/dna-microarray-support.html)). L'analyse des

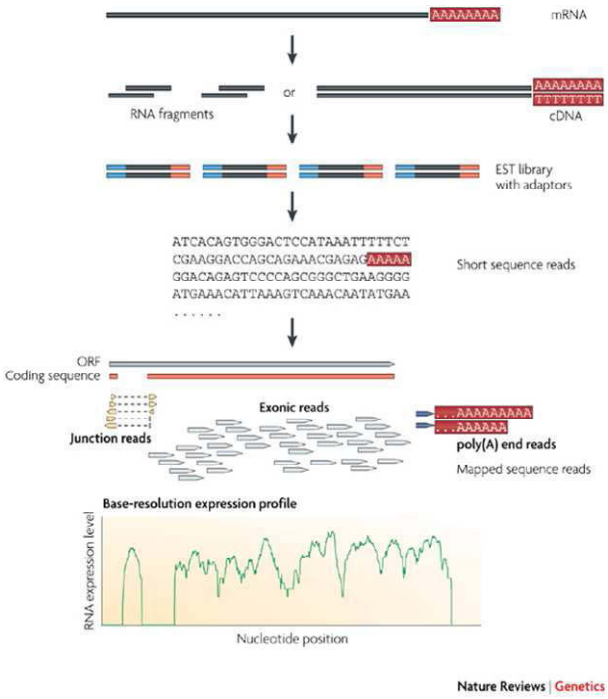



**Figure 21** : Illustration d'une expérience RNA-Seq.
Les ARN sont sélectionnés grâce à leur queue polyA, puis fragmentés et reverses transcrits, ou reverses transcrits en ADNc puis fragmentés. Les adaptateurs (bleu) sont ajoutés aux ADNc, qui sont séquencés par la suite. Il en résulte de courtes lectures qui peuvent être alignées sur une séquence de référence, pour construire au final le profil d'expression de chaque gène. D'après (Z. Wang, Gerstein, et Snyder 2009).

**Tableau 3** : Comparaison de trois méthodes d'analyse du transcriptome.
D'après (Z. Wang, Gerstein, et Snyder 2009).

| Technology | Tiling microarray | cDNA or EST sequencing | RNA-Seq |
|---|---|---|---|
| *Technology specifications* | | | |
| Principle | Hybridization | Sanger sequencing | High-throughput sequencing |
| Resolution | From several to 100 bp | Single base | Single base |
| Throughput | High | Low | High |
| Reliance on genomic sequence | Yes | No | In some cases |
| Background noise | High | Low | Low |
| *Application* | | | |
| Simultaneously map transcribed regions and gene expression | Yes | Limited for gene expression | Yes |
| Dynamic range to quantify gene expression level | Up to a few-hundredfold | Not practical | >8,000-fold |
| Ability to distinguish different isoforms | Limited | Yes | Yes |
| Ability to distinguish allelic expression | Limited | Yes | Yes |
| *Practical issues* | | | |
| Required amount of RNA | High | High | Low |
| Cost for mapping transcriptomes of large genomes | High | High | Relatively low |

profils d'expression par l'utilisation des puces à ADN reste une bonne alternative à une approche par séquençage. En 2011, 520 000 expériences utilisant des puces à ADN ont été recensé sur le site GEO (« Gene Expression Omnibus ») (Malone & Oliver, 2011).

Cependant, ces méthodes présentent des limites, soit dans leur capacité à détecter de nouveaux transcrits, soit à quantifier l'expression ou encore, à étudier le transcriptome dans sa globalité. Effectivement, ce sont des méthodes dites « figées », du fait de l'utilisation de sondes/amorces, qui implique la connaissance d'une séquence de référence. De plus, la mesure du niveau d'expression par l'utilisation de puce à ADN peut aussi être entachée par du bruit de fond, du à une hybridation non spécifique d'ADNc, partiellement complémentaire des sondes.

## 3.3 Le RNA-Seq : un outil haut débit pour l'analyse du transcriptome

La méthode récente basée sur le séquençage haut débit, appelée RNA-Seq pour « RNA sequencing » , a permis de lever les multiples verrous mentionnés dans le paragraphe précédent. Le RNA-Seq consiste à isoler les ARNm d'une cellule ou d'un tissu, puis de les fragmenter pour ensuite les réverse complémenter afin d'obtenir des fragments d'ADNc, qui seront par la suite séquencés (Figure 21) (Ozsolak & Milos, 2011). Deux approches sont possibles pour le traitement des lectures : l'assemblage *de novo* ou bien l'alignement (« mapping »).

Le RNA-Seq est une approche particulièrement efficace pour étudier des organismes pour lesquels aucun génome de référence n'est disponible. Dans la mesure où c'est une approche qui permet de générer un transcriptome de référence moyennant une étape d'assemblage *de novo* des lectures. Le RNA-Seq présente aussi l'avantage de fournir des mesures précises du niveau d'expression des gènes ainsi que de chacun des transcrits alternatifs. Par ailleurs, le polymorphisme nucléotidique entre gènes dupliqués (notamment chez les polyploïdes) peut être pris en compte pour mesurer l'expression spécifique d'une copie. Enfin, il permet de mettre en évidence de nouvelles régions transcrites non annotées.

La première étude basée sur le RNA-Seq a été publiée en 2006 (Bainbridge et al., 2006) (sur cellules humaines cancéreuses). En 2009, une étude a été publiée montrant le RNA-Seq comme un outil révolutionnaire pour les analyses de transcriptomique, en termes de résolution de détection des transcrits, d'applications pour la détection de transcrits et de l'expression allèle spécifique, de quantité de matériel biologique utilisé et aussi de coût (Tableau 3) (Z. Wang, Gerstein, & Snyder, 2009). Le nombre de lectures générées est un paramètre crucial qui définit la précision avec laquelle il est possible d'étudier le niveau

| Package | Input | Category | Uses |
|---|---|---|---|
| **Read Mapping** | | | |
| BFAST | Reference Genome or reference transcriptome Sequence reads | Unspliced seed aligners | Expression quantification of a known set of transcripts with reads from low quality data or of species with high mutation rates that may contain indels at high frequency SNP discovery in expressed transcripts Long reads (e.g. from Roche 454) |
| SHRiMP | | | |
| Stampy | | | |
| Bowtie | Burrows-Wheeler transformed reference genome or transcriptome | BWT short read aligners | Alignment to an existing transcriptome for quantification. Some spliced-read aligners build on top of these fast aligners to map unspliced reads or read segments. |
| BWA | | | |
| SOAP2 | | | |
| MapSplice | Burrows-Wheeler transformed reference genome or transcriptome | Exon first spliced aligners | The sensitivity, speed and relatively low resources required by these aligners makes them ideal for genome guided transcript reconstruction and expression quantification methods. |
| SpliceMap | | | |
| TopHat | | | |
| **Transcript reconstruction** | | | |
| Scripture | Aligned reads to a reference genome | Genome-guided transcript reconstruction methods | Transcript reconstruction, expression quantification. Ideal to find novel, previously unannotated transcripts. |
| Cufflinks | | | Transcript reconstruction, gene/transcript expression quantification and differential analysis of expression alternative splicing. |
| Velvet (OASES) | Sequence reads | Whole-genome assemblers | Reconstructing genomes from short-read data at high coverage. These methods can be used in principle for transcriptomes but they are not optimized to do so. |
| Trans- ABySS | Sequence reads | Transcriptome assembler | Reconstructing transcripts including alternative isoforms. This method is optimized to reconstruct a transcriptome without using a known genome reference. |
| **Transcript quantification** | | | |
| Alexa-Seq | Reference or reconstructed | Gene quantification | Find differentilly included exons. |
| ERANGE | | | Quantify gene expression. |
| Cufflinks | Reference or reconstructed | Transcript quantification | Quantify transcript expression, differential isoform expression and alternative splicing analysis. |
| MISO | | | |
| Cuffdiff | Reference or reconstructed | Differential Expression | Find differentially expressed or spliced genes |
| DegSeq | | | Find differentially expressed genes |
| EdgeR | Pre-computed fragment counts for each gene and | | |
| DESeq | | | |
| Myrna | Sequence reads | | |

**Tableau 4** : Liste non exhaustive des outils d'analyse de données RNA-Seq.
D'après (Garber et al. 2011)

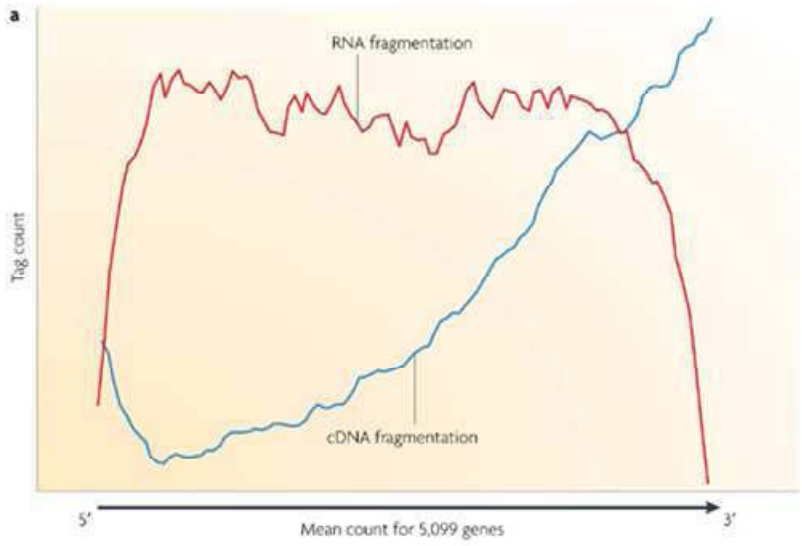



**Figure 22 :** Influence de la préparation des librairies sur la couverture des transcrits.
La fragmentation de l'ADNc (bleu) entraine un biais en faveur de l'extrémité 3' du transcrit. La fragmentation de l'ARN (rouge) entraine une couverture plus important au niveau du corps du transcrit, mais une déplétion au niveau des extrémités 5' et 3'. L'axe y représente la couverture moyenne des lectures pour 5000 ORF de levure. D'après (Z. Wang, Gerstein, et Snyder 2009).

d'expression de l'ensemble des gènes. Il a été montré que 30 et 20 millions de lectures sont suffisantes pour couvrir, respectivement, le transcriptome d'un individu et d'une cellule, à partir d'une étude sur différents phylums (Annélides, Arthropodes, Chordés, Cnidaires, Cténophores et Mollusques) (Francis et al., 2013). Cette étude précise aussi qu'au delà de 60 millions de lectures, seulement un très faible nombre de nouveaux gènes peut être mis en évidence, alors qu'il y a une forte accumulation d'erreurs de séquençage pour les gènes fortement exprimés (Francis et al., 2013).

Le RNA-Seq et l'hybridation sur puces à ADN restent néanmoins des techniques complémentaires. Une comparaison de la performance entre les deux technologies en termes de quantification du niveau d'expression des gènes, en faisant un focus sur plusieurs aspects comme la reproductibilité, la précision, les variabilités biologiques et techniques, ainsi que les problèmes techniques et statistiques de l'analyse a montré que : malgré une forte correspondance entre les niveaux d'expression mesurés dans les deux approches, tous les autres paramètres plaident en faveur du RNA-Seq (meilleures reproductibilité et précision notamment) (Kogenaru, Qing, Guo, & Wang, 2012). La quantification du niveau d'expression des gènes faiblement exprimés est particulièrement mieux appréhendée avec le RNA-Seq (Zhao, Fung-Leung, Bittner, Ngo, & Liu, 2014 ; Y. Guo et al., 2013). Cette approche comporte néanmoins de nombreux défis. Le premier repose sur la construction d'une banque de fragments d'ADNc non biaisée. En effet, lors de la construction de la librairie RNA-Seq, la fragmentation de l'ADNc entraine un biais en faveur de la partie 3' des transcrits alors que la fragmentation de l'ARNm favorise une meilleur couverture du corps des transcrits, mais une déplétion au niveau des extrémités 5' et 3' (Z. Wang et al., 2009) (Figure 22). De plus, l'amplification des ADNc avant séquençage peut entraîner des bien pour la quantification de l'expression des gènes. Le second défi réside dans l'analyse bioinformatique des données. Bien qu'il existe des centaines d'outils d'analyse de données RNA-Seq (Garber, Grabherr, Guttman, & Trapnell, 2011) (Tableau 4), leur mise en œuvre n'est pas encore devenue une routine pour tous les laboratoires. Un inconvénient majeur est le volume des données numériques générées, qui pose des problèmes de stockage informatique et de puissance de calcul.

L'approche RNA-Seq a été utilisée pour analyser de nombreux transcriptomes de plantes, comme par exemple : la pêche (L. Wang et al., 2013), *A. thaliana* (Giorgi, Fabbro, & Licausi, 2013), le maïs (Hansey et al., 2012), le prunus (Martínez-Gómez, Crisosto, Bonghi, & Rubio, 2011), la vigne (Zenoni et al., 2010), B. *distachyon* (International & Initiative, 2010).

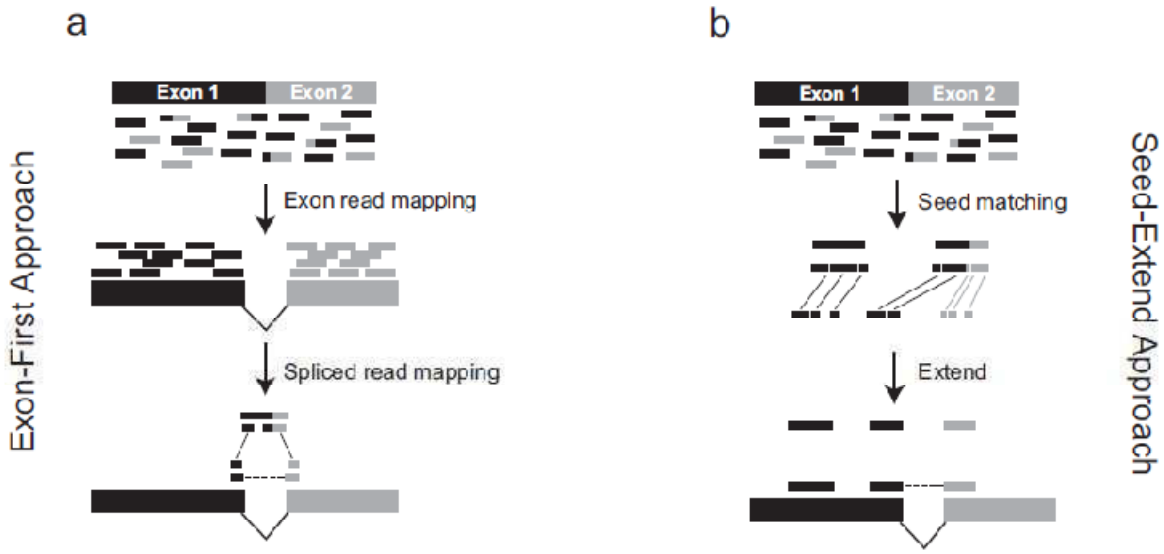



**Figure 23** : Méthodes d'alignement des lectures issues de RNA-Seq, sur une séquence de référence. Les cours noire et grise représentent l'origine des différentes lecture (a) : Approche « Exon-First » où les lectures non épissées sont alignées en premier. Puis les lectures qui n'ont pas été alignées lors du premier tour sont alignées au niveau des jonctions d'épissage. (b) Approche « Seed-Extend », chaque lecture est divisée en k-mer, ces derniers sont alignés sur la séquence de référence, puis l'alignement est étendu en prenant en compte les trous formés par les sites d'épissage. D'après (Garber et al. 2011).

### 3.3.1 Analyse des lectures issues du séquençage

La première étape de la partie de bioanalyse des données RNA-Seq consiste à vérifier la qualité des lectures issues de séquençage. Avec la technologie Illumina, les lectures ont une taille généralement comprise entre 75 à 300 nucléotides en fonction de la technologie de séquençage utilisée. Chaque nucléotide est associé à un signal colorimétrique transformé en score de qualité (noté Q) correspondant à une probabilité d'erreur de séquençage (Ewing & Green, 1998). Généralement, la valeur seuil pour Q est fixée à 30, représentant une probabilité d'erreur de 1/1000. Pour des données issues de séquençage Illumina, les erreurs de séquençage ne sont pas distribuées aléatoirement sur les lectures, et augmentent dans les extrémités 5' et 3' (B. Liu et al., 2012).

L'approche standard pour faire face aux nucléotides de mauvaise qualité est d'éliminer les régions/nucléotides de mauvaise qualité (« trimming »). Il existe plusieurs outils pour réaliser cette étape : FASTX quality trimmer, PRINSEQ, Trimmomatic, ConDe Tri (« FASTX-Toolkit »; Lohse et al., 2012; Schmieder & Edwards, 2011; Smeds & Künstner, 2011). La comparaison de différents outils de « trimming », ainsi que les seuils de Q, pour des analyses RNA-Seq (assemblage *de novo*, génotypage) montre que l'interprétation des données peut être influencée par la combinaison du seuil de Q ainsi que de l'outil de « trimming » utilisés (Del Fabbro, Scalabrin, Morgante, & Giorgi, 2013). De plus, si généralement des seuils assez drastiques de Q (ex : Q=20) sont utilisés, des valeurs de score Phred de 2 ou 5 sont plus optimales pour la plupart des analyses de données (Macmanes, 2014). Une autre approche est de ne pas réaliser cette étape, car les lectures de mauvaise qualité ne seront pas alignées si des paramètres non permissifs sont utilisés pour l'alignement des lectures.

Lorsqu'une séquence de référence est disponible, un grand nombre d'outil est disponible pour l'alignement des lectures (Garber et al., 2011). L'alignement des lectures RNA-Seq, à la différence de l'alignement de lectures d'ADNg, nécessite des outils adaptés pour la gestion des trous formés par les introns. Pour cela des outils, dits « spliced-aligners » ont été développés, et sont séparés en deux catégories : « exon-first » et « seed-and extend » (Figure 23). Ces deux méthodes se distinguent par la gestion de l'alignement des lectures aux jonctions intron/exon. Dans la méthode « seed-and-extend », les lectures sont découpées en courtes régions (« seed ») puis, ces régions sont alignées sur le génome. Alors qu'avec la méthode « exon-first », l'alignement se fait en deux étapes. Dans un premier temps les lectures sont alignées sans recherche des sites d'épissage, puis les lectures non alignées sont utilisées pour identifier précisément les jonctions intron/exon. Les études

portant sur la comparaison des outils et des approches ont montré que l'efficacité des outils bioinformatiques est fonction du taux d'erreur de séquençage, du nombre de lectures produites, du temps d'analyse et de la complexité du génome étudié (Hatem, Bozdağ, Toland, & Çatalyürek, 2013; Soneson & Delorenzi, 2013).

### 3.3.2 Calcul du niveau d'expression/normalisation des données et expression différentielle

La quantification du niveau d'expression des gènes nécessite une étape de reconstruction des transcrits. Des outils tels que Cufflinks (identifie le nombre minimal d'isoformes ; (Trapnell et al., 2010)) ou bien Scripture (identifie toutes les isoformes ; (Guttman et al., 2010)) permettent de réaliser cette étape.

Le niveau d'expression est fonction du nombre de lectures alignées et nécessite une étape clé de normalisation qui tient compte à la fois de la fragmentation des ARNm ou ADNc lors la construction de la banque, de la taille des ARNm, et du nombre de lectures produites (Garber et al., 2011, p. 2). Les indices RPKM/FPKM (« Reads/Fragments Per Kilobase of transcript per Million mapped reads ») constituent une méthode de normalisation de l'expression des transcrits, en prenant en compte le nombre de lectures générées respectivement à partir d'une (« single ends ») ou des deux extrémités (« paired-ends ») de chaque fragment d'ADNc pour un transcrit, ainsi que le taille du transcrit (Mortazavi, Williams, Mccue, Schaeffer, & Wold, 2008). Certaines lectures s'alignent à plusieurs endroits sur le génome, affectant ainsi la quantification de l'expression. Les outils comme Cufflinks et MISO (Katz, Wang, Airoldi, & Burge, 2010) permettent de gérer cette incertitude par la construction d'une fonction de vraisemblance qui modélise le processus de séquençage et identifie les estimations de l'abondance des isoformes (Trapnell et al., 2010). L'expression d'un gène est définie par la somme de l'expression de ses isoformes. Cependant, le calcul de l'expression des différentes isoformes est une étape critique. Deux méthodes sont utilisées : la méthode « *intersection d'exons* » qui compte les lectures qui s'alignent sur les exons constitutifs (Bullard, Purdom, Hansen, & Dudoit, 2010), et la méthode « *union d'exons* » qui compte toutes les lectures qui s'alignent sur chaque exon, pour toutes les isoformes d'un gène (Griffith et al., 2010; Mortazavi et al., 2008).La méthode « *union d'exons* » semble surestimer l'expression des gènes avec épissage alternatif alors que la méthode « *intersection d'exons* » réduit la puissance d'analyse d'expression différentielle (Garber et al., 2011).

L'expression des gènes n'est pas uniforme pour toutes les conditions ou tissus d'une analyse RNA-Seq, et afin de détecter les changements d'expression des gènes entre les

**Tableau 5** : Différentes espèces de blé (genres *Aegilops*, *Amblyopyrum* et *Triticum*).
D'après (Feldman et Levy 2012)

| Ploidy level | Species | Genome |
|---|---|---|
| Diploids | *Amblyopyrum muticum (=Ae. mutica)* | TT SS |
| (2*n* = 2*x* = 14) | *Aegilops speltoides Ae. bicornis* | SbSb SlSl SlSl SsSs DD |
| | *Ag. Longissima Ae. sharonensis Ae. Searsii* | CC UU MM NN |
| | *Ae. tauschii* | AmAm AA |
| | *(=Ae. squarrosa) Ae. caudata* | |
| | *Ae. umbellulata Ae comosa* | |
| | *Ae. uniaristata Triticum monococcum* | |
| | *T. urartu* | |
| Tetraploids | *Ae. biuncialis Ae. eniculata(=Ae. ovata)* | UUMM |
| (2*n* = 4*x* = 28) | *Ae. neglecta (=Ae. triaristata 4x)* | MMUU |
| | *Ae*, *columnaris Ae. triuncialis Ae. kotschyi Ae. peregrina* | UUMM |
| | *(=Ae. variabilis) Ae. cylindrica* | UUMM UUCC; CCUU SSUU |
| | *Ae. crassa* 4x Ae. Ventricosa | SSUU |
| | *T. turgidum T. timopheevii* | DDCC DDMM DDNN |
| | | BBAA GGAA |
| Hexaploids | *Ae. recta (=Ae. triaristata 6x)* | UUMMNN |
| (2*n* = 6*x* = 42) | *Ae. vavilovii Ae. crassa* 6x Ae. juvenalis | DDMMSS DDDDMM DDMMUU |
| | T. aestivum T, zhukovskyi | BBAADD GGAAAmAm |

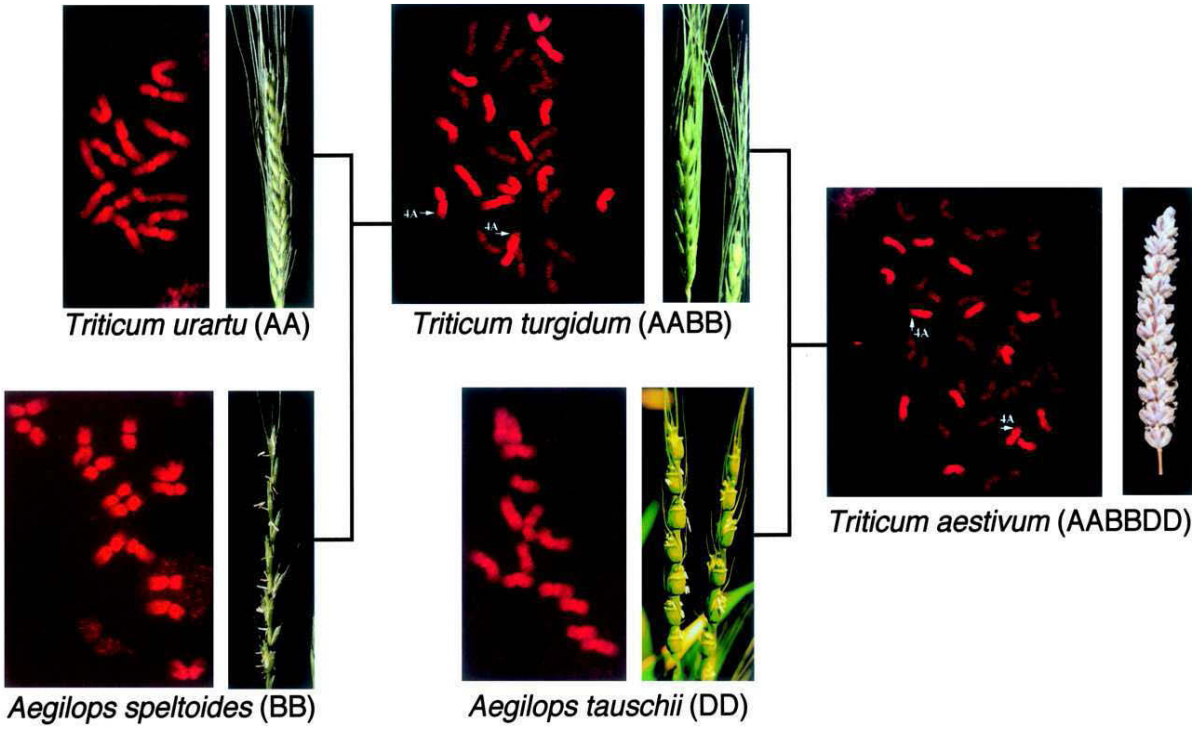



**Figure 24** : Schéma des deux évènements de polyploïdisation qui ont conduit à la formation du blé hexaploïde.
D'après (B. S. Gill et al. 2004).

conditions étudiées, il faut passer par une étape d'analyse d'expression différentielle. Les méthodes de calcul se divisent en deux catégories : les modèles paramétriques et les modèles non paramétriques. Les approches paramétriques se basent sur une distribution des probabilités connues comme par exemple les lois Binomiale, de Poisson ou bien Binomiale Négative, alors que les approches non paramétriques n'ont pas d'apriori sur la distribution des données. Le choix de la méthode de normalisation va donc avoir un impacte sur la détection des gènes différentiellement exprimés.

En conclusion, le RNA-Seq est une approche très puissante pour l'analyse du transcriptome. Elle permet de détecter des niveaux d'expression même très faibles. De plus, les transcrits mis en évidence peuvent être utilisés pour l'annotation structurale des gènes. Cependant, elle comporte de nombreux défis, notamment concernant les approches statistiques. La construction des banques est une étape qui impacte les conclusions tirées de l'analyse, et la normalisation influence considérablement les résultats d'expression différentielle.

# 4 Le génome du blé tendre

## 4.1 Origine du génome hexaploïde

Le blé tendre *Triticum aestivum L.* est une plante monocotylédone de la famille des *Poaceae* (graminées). C'est la deuxième source de calories pour l'Homme après le riz (FAOstat). Sa domestication remonterait à près de 9000 ans au proche Orient (Shewry, 2009). Au delà du blé tendre, le groupe des blés appartenant aux genres *Aegilops* et *Triticum* est constitué de 13 espèces diploïdes et 18 espèces allopolyploïdes (Feldman & Levy, 2012) (Tableau 5). Ce groupe constitue un modèle d'étude de l'évolution des génomes polyploïdes (Feldman & Levy, 2012).

Le blé tendre est une espèce allohexaploïde issue de deux événements récents d'hybridation. Le premier événement s'est produit il y a environ 500 000 ans (B. S. Gill et al., 2004) entre deux espèces dont les plus proches représentants actuels sont *Triticum urartu* (génome AA) et *Aegilops speltoides* (génome SS), donneur du génome B. Le tétraploïde issu de cet événement est *Triticum dicoccoïdes*, ancêtre du blé dur *Triticum turgidurum* (AABB, 2n=14). Il y a environ 10 000 ans, une seconde hybridation avec *Aegilops tauschii* (génome DD), a donné naissance au blé tendre (Figure 24). Son génome regroupe donc 3 sous-génomes homéologues diploïdes à 7 paires de chromosomes chacun. La divergence des génomes A, B et D a été estimée entre 2,5 et 4,5 millions d'années (Shaoxing Huang et al., 2002). Ainsi, les chromosomes homéologues partagent un contenu en gènes très similaire avec un niveau d'identité élevé d'en moyenne 97% (International Wheat Genome

Sequencing Consortium, 2014). Toutefois, le contenu en éléments transposables, qui composent la majeure partie du génome et qui, pour la plupart, ne sont pas soumis à pression de sélection, a été largement réarrangé au cours des 2,5-4,5 millions d'années. Par exemple, la comparaison de trois régions homéologues porteuses du locus Ha (contrôlant la dureté du grain) a montré une absence complète de conservation des séquences intergéniques (Chantret et al., 2005).

La taille des sous-génomes AA, BB et DD est proche. Le contenu en ADN de *T. urartu*, *Ae. tauschii* et *Ae. speltoides* a été estimée à 6,02, 5,17 et 5,81 pg, respectivement, ce qui représente une taille moyenne de 5,5 Gb avec la relation 1 pg=978 Mb (Eilam et al., 2007). Le génome du blé tendre compte environ 17 Gb, ce qui représente plus de 7 fois la taille du génome du maïs (2,3 Gb). Dans ce génome de grande taille, seulement 2 à 3% de la séquence est codante, et 80 à 90% sont des séquences dérivées d'éléments transposables. Bien que le blé soit hexaploïde, il se comporte comme un diploïde en méiose, du fait de l'absence d'appariement entre chromosomes homoéologues (par exemple : les chromosomes 1A et 1B). Ce phénomène est contrôlé par un locus qui empêche la formation de bivalents entre les chromosomes homéologues : le locus Ph1. Ce locus est situé sur le chromosome 5B, ainsi que sur ses homéologues 5A et 5D mais n'est pas actif sur ces derniers (Griffiths et al., 2006). Si ce locus est inactivé, il y a formation de multivalents entre les chromosomes homéologues pendant la première division de méiose et appariement de chromosomes au sein d'hybrides de blé (Vega & Feldman, 1998). Le rôle de Ph1 serait d'induire une modification de la structure des chromosomes de manière synchronisée entre les deux homologues pour permettre leur reconnaissance, leur appariement, la formation des crossing-overs et ainsi leur ségrégation correcte (Martinez-Perez, Shaw, & Moore, 2001).

## 4.2 Stratégies de séquençage du génome

### 4.2.1 Impact de la taille du génome sur la stratégie d'assemblage

L'augmentation du nombre de génomes de plantes séquencés suit le rythme de l'augmentation des débits des technologies de séquençage (Michael & Jackson, 2013). Les technologies actuelles permettent de séquencer des milliards de bases par jour pour un coût faible (Zhou et al., 2010). Toutefois, si la production de lectures n'est généralement plus le facteur limitant, l'assemblage des données reste un problème. Plus les lectures sont courtes, plus l'assemblage donnera des contigs de petite taille. Ainsi, l'assemblage des lectures produites pour des génomes de grande taille (>1 Gb) via des approches génome-entier (« whole genome shotgun ») aboutit à des millions de contigs de petite taille, plutôt qu'à une

séquence unique par chromosome. Comme mentionné dans le paragraphe 1.1, les génomes de plantes sont parfois cent fois plus grands que les génomes des mammifères, poissons ou oiseaux actuellement séquencés. De plus, ces génomes sont souvent polyploïdes et parfois hétérozygotes. Par ailleurs, la présence de grandes familles de gènes répétés, l'abondance de pseudogènes et de séquences issues de duplications récentes, l'activité des transposons, ainsi que des insertions multiples de génomes mitochondriaux et chloroplastiques dans le génome nucléaire démultiplient les problèmes liés à l'assemblage *de novo* (Schatz et al., 2012). Pour toutes ces raisons, le séquençage et l'assemblage *de novo* des génomes de plantes entrainent des résultats fragmentés. Le pendant de cette fragmentation, est que plus l'assemblage d'un génome est fragmenté, plus l'utilisation de la séquence pour répondre à des questions biologiques sera limitée.

Produire une séquence de référence d'un génome ne se limite donc pas à produire suffisamment de lectures, mais bien à établir une stratégie la plus adaptée à l'obtention d'une séquence assemblée utile. Le modèle de Lander-Waterman (Lander & Waterman, 1988) offre une prédiction analytique de la couverture de séquençage minimum requise pour assembler de grands contigs. En utilisant ce modèle, une couverture minimum de 15X est nécessaire pour assembler des lectures de 100 pb en grands contigs pour les génomes de l'homme et de la sourie (Gnerre et al., 2011). Cependant, la couverture est à optimiser en fonction de la taille du génome, du niveau de ploïdie, du taux de séquences répétées, et du taux d'erreurs de séquençage. Pour les génomes complexes (>1 Gb), une couverture de 100X est recommandée si les lectures sont courtes (100 pb) (Gnerre et al., 2011).

### 4.2.2 Trouver la meilleure stratégie pour simplifier l'assemblage

Le premier génome séquencé a été celui d'*A. thaliana* en 2000 (Arabidopsis Genome Initiative, 2000). Depuis, un grand nombre de génomes de plante a été séquencé (Michael & Jackson, 2013). Comme par exemple : le riz (Goff et al., 2002), le peuplier (Tuskan et al., 2006), la papaye (Ming et al., 2008), le sorgho (Paterson et al., 2009b), le soja (Schmutz et al., 2010), la pomme de terre (Xu et al., 2011), l'orge (International Barley Genome Sequencing Consortium et al., 2012), la tomate (The Tomato Genome Consortium, 2012) et plus récemment la banane (D'Hont et al., 2012).

On peut distinguer deux approches majeures pour le séquençage de ces génomes :

1/ le séquençage via la production d'une banque de grands fragments ordonnés (c'est à dire une carte physique; à partir de BAC, YAC).

2/ le séquençage via la production d'une librairie de fragments issus du génome complet (WGS), c'est-à-dire sans étape de sous-clonage de fragments de grande taille. Les débits très élevés des nouvelles technologies de séquençage (NGS) ont rendu possible la production, à moindre coût, d'un nombre de lectures suffisamment élevé pour atteindre des niveaux de couverture satisfaisants (>20X) pour n'importe quel génome. C'est la raison pour laquelle les projets de séquençage de ces dernières années se sont affranchis de l'étape de création d'une carte physique, coûteuse en temps et en argent. Cette méthode peut aussi être adaptée au séquençage d'un chromosome (WCS).

Les technologies de séquençage offrent la possibilité de séquencer différents types de librairies dites « mate pair », « paired end » et « single-end », avec une taille variable de plus de 100 pb (jusqu'à 300 pb avec la technologie MiSeq). Une des promesses des technologies de séquençage troisième génération est de générer des lectures de grande taille comme par exemple la technologie développée par la société Pacific Bioscience et celle en cours de mise au point par la société Oxford Nanopore Technology.

Les premiers génomes séquencés ont été ceux d'Arabidopsis (150 Mb) (Arabidopsis Genome Initiative, 2000) et du riz (400 Mb) (Goff et al., 2002). Pour ces deux génomes, c'est l'approche par séquençage de BAC qui a été utilisée. Les BAC qui ont été sélectionnés à partir d'une carte physique, représentent le plus court chemin pour couvrir l'ensemble du génome (MTP; Minimum Tilling Path : clones couvrant une région/contig avec le minimum de redondance), et le séquençage a été réalisé avec la technologie Sanger. La démonstration de l'utilisation de l'approche de séquençage WGS a été faite avec le séquençage des génomes de la souris (Mouse Genome Sequencing Consortium et al., 2002) et de l'homme (Venter et al., 2001), et ce en montrant qu'il n'était pas nécessaire d'utiliser une carte physique, ce qui a stimulé l'engouement pour cette approche pour une grande partie des génomes de plantes (Feuillet, Leach, Rogers, Schnable, & Eversole, 2011). Cependant en comparant les approches WGS et BAC par BAC pour le génome de la souris, l'étude a montré que 5% de la séquence du génome (267 Mb) étaient assemblés de façon incorrecte ou manquaient dans l'approche WGS (Church et al., 2009). Et à l'exception du maïs et du riz, l'approche par WGS a été utilisée pour tous les génomes de plantes par exemple : *Sorghum bicolor* (Paterson et al., 2009b), *Vitis vinifera* (Jaillon et al., 2007), *B. distachyon* (International Brachypodium Initiative, 2010), *Glycine max* (Schmutz et al., 2010), *Populus trichocarpa* (Tuskan et al., 2006), *Cucumis sativus* (Sanwen Huang et al., 2009).

Le challenge pour le séquençage des génomes complexes réside dans le contournement de deux caractéristiques réduisant l'efficacité de l'assemblage : la fraction répétée du génome (notamment les éléments transposables et les séquences satellites) et le niveau de ploïdie.

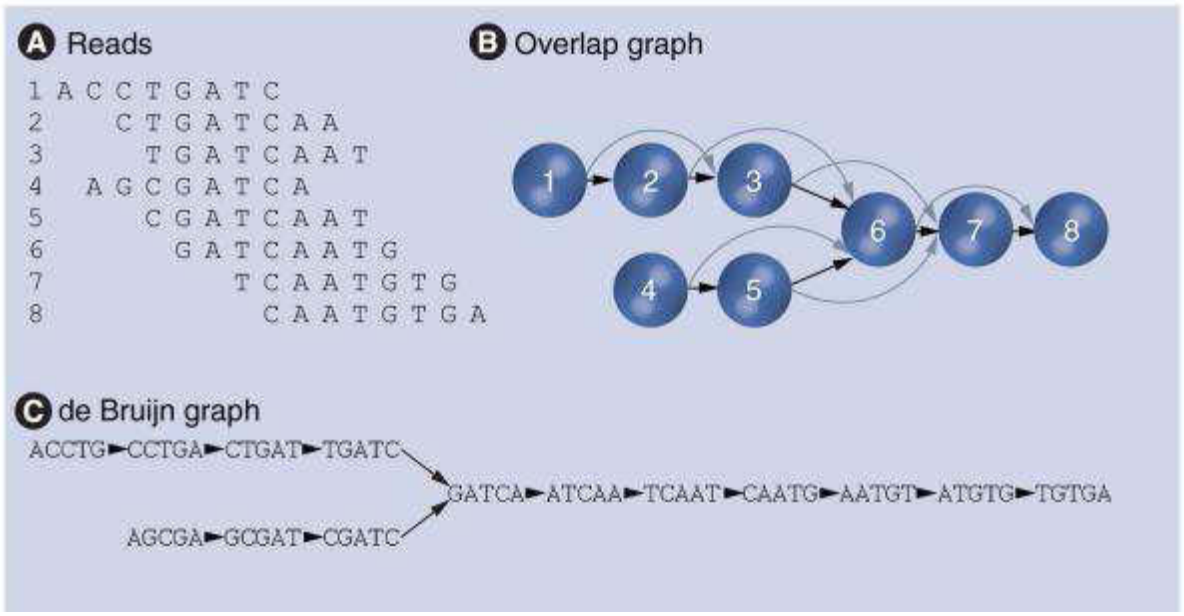



**Figure 25** : Construction d'une graphe pour l'assemblage de lectures
(A) Représentation des 8 lectures à aligner.
(B) Graphe de chevauchement correspondant, les ronds correspondent aux lectures et les flèches aux chevauchements de 5 nucléotides au moins.
(C) Construction du graphe de Bruijn.

La stratégie BAC par BAC permet de s'affranchir du caractère répété d'une séquence car une répétition à l'échelle du génome complet peut être une séquence unique au sein d'un BAC. Une autre possibilité pour réduire la complexité des génomes est le tri des chromosomes par cytométrie en flux. C'est la stratégie qui a été retenue pour produire une séquence de référence du génome du blé hexaploïde par le Consortium International du Séquençage du Génome du Blé (« IWGSC », www.wheatgenome.org).

### 4.2.3 Assemblage, « scaffolding » et ancrage des séquences le long des chromosomes

L'assemblage d'un génome consiste généralement en une multitude de contig (séquences contigües) et de scaffolds (groupes de contigs séparés par une distance connue) couvrant environ 90% ou plus de la taille du génome en question (Chain et al., 2009). Traditionnellement, les contigs sont construits à partir de lecture « paired-end », ayant une distance inférieure à 3 kb entre les deux extrémités de chaque fragment. Pour construire les scaffolds, ce sont les lectures « mate paire » qui sont utilisées, avec une distance supérieure à 3 kb. Cette étape est appelée scaffolding, et permet d'ordonner et d'orienter les contigs précédemment construits. Généralement, l'étape d'assemblage suit un processus hiérarchique, en comparant les lectures pour former un graphique ou chemin d'assemblage des lectures qui se chevauchent sur des k-mer. Les k-mer sont des motifs (ou mots) d'une longueur k observés plus d'une fois dans une séquence génomique. Le but est de créer le chemin le plus simple pour former des contigs. Traditionnellement, l'assemblage des lectures de séquençage passe par la construction de structure de graphe, les plus courant étant le graphe de Bruijn (Figure 25), et « overlap–layout–consensus » (OLC) (Henson, Tischler, & Ning, 2012). La qualité de l'assemblage est le reflet de la contigüité (l'inverse de la fragmentation), qui peut être estimée par de nombreux critères. Les valeurs statistiques les plus couramment utilisées sont la N50 et la « N50 length » des contigs et des scaffolds. Considérant un set de contigs/scaffolds issu d'un assemblage de génome, la « N50 length » correspond à la taille pour laquelle les contigs/scaffolds plus grands que cette taille représentent 50% de la séquence assemblée. La N50 définit le nombre de contigs/scaffolds qui ont une taille supérieure ou égale à la « N50 length ». Plus la « N50 length » est grande, plus la contiguité de l'assemblage est élevée (Yandell & Ence, 2012). Yandell (2012) propose qu'un assemblage puisse être utilisé pour une annotation, si la « N50 length » d'un scaffold est équivalente à la taille d'un gène. Si la « N50 length » des scaffolds correspond à la taille médiane des gènes, alors 50% de ces gènes sont assemblés en un seul scaffold et 50% sont fragmentés dans l'assemblage.

### 4.2.4 Application au génome du blé

Le blé est la dernière céréale majeure pour laquelle aucune séquence de référence complète est disponible (Feuillet et al., 2011). Afin de fournir cette ressource manquante, plusieurs approches utilisant différentes stratégies ont été mises en place.

Les génomes de *T. urartu* et *Ae. tauschii* ont été séquencés en 2013 via une approche génome entier (« Whole Genome Shotgun », WGS) sur une plateforme Illumina (Jia et al., 2013; Ling et al., 2013). L'assemblage du génome de *T. urartu* a généré 81 689 scaffolds d'une taille supérieure à 2 kb, représentant 4,66 Gb au total. Les éléments transposables représentent 67% des séquences assemblées, dont les trois quarts sont des rétrotransposons à LTR. La même stratégie de séquençage a été utilisée pour *Ae. tauschii* (Jia et al., 2013), à partir de 45 banques de fragments d'ADN allant de 0,2 à 20 kb séquencés par les deux extrémités. L'assemblage a généré 6 995 685 scaffolds représentants 4,23 Gb. De plus, ils ont montré que 66% du génome est composé d'éléments transposables regroupés en 410 familles différentes parmi lesquelles ont en retrouve 20 majoritaires qui à elles seules représentent plus de 50% du génome. Cependant, ce pourcentage est sous-estimé, par rapport aux 80% attendu, la raison vient de la stratégie de séquençage utilisée. Une approche similaire a été utilisée pour le séquençage du blé hexaploïde. Ainsi, en 2012, près de 100 Gb de lectures Roche/454 ont été produites à partir de l'ADN du génome entier de la variété Chinese Spring (qui est la variété de référence internationale), représentant une couverture de 5X (Brenchley et al., 2012). Cependant, cette ressource ne correspond pas à une séquence de référence. Les lectures ont été assemblées par similarité avec les séquences de gènes orthologues, ce qui ne reflète pas le génome dans son ensemble.

Afin de produire une séquence de haute qualité à la communauté scientifique et aux sélectionneurs, le Consortium IWGSC a été créé en 2005. La stratégie établie par le Consortium est fondée sur la réduction de la complexité du génome hexaploïde via le tri de chromosome par cytométrie de flux (Kubaláková et al., 2005). La vision du Consortium IWGSC est de ne pas aborder le génome dans son intégralité, mais bien de réduire sa complexité via, tout d'abord, le tri de chromosome, puis, la construction de banques BAC et l'établissement de cartes physiques pour les différents chromosomes. La stratégie pour l'assemblage d'une séquence de haute qualité passe donc par le séquençage des BAC ordonnés le long des chromosomes, avec comme chromosome témoin le chromosome 3B (Choulet et al., 2014). Toutefois, sans attendre la fin de la construction des cartes physiques de tous les chromosomes, le Consortium a décidé de profiter de l'augmentation des débits

de séquençage (des plateformes Illumina notamment) pour produire une ressource de qualité intermédiaire : un assemblage « shotgun » de chacun des bras de chromosomes triés en cytométrie de flux. L'idée était ici d'établir un assemblage, certes fragmenté, mais assignable à chacun des chromosomes. Cette ressource a été nommée « Chromosome Survey Sequences » (CSS). Chaque bras de chromosome (hormis le 3B trié entier) a été séquencé avec la technologie Illumina à une couverture de 30X à 234X (paired-ends 2x100 pb). Les lectures ont été assemblées en contigs représentant 10,2 Gb au total avec une valeur N50 de 2,2 kb en moyenne. Le contenu en éléments transposable des séquences a ainsi été estimé à 77% (International Wheat Genome Sequencing Consortium, 2014).

## 4.3 Composition et organisation du génome

### 4.3.1 Estimation du nombre de gènes

Le contenu en gènes du génome de blé reste aujourd'hui une estimation du fait de l'absence d'une séquence de référence, et peut se faire soit sur la base de la conservation des gènes entre les espèces (Devos et al., 2005; Keller & Feuillet, 2000), soit à partir d'extrapolations de l'annotation de l'assemblage de séquençage de BAC. A partir d'échantillons de séquences produites à l'échelle du génome complet, en 2005, Rabinowicz a estimé le nombre de gènes par sous-génome à 98 640, soit au total 295 900 gènes (Rabinowicz et al., 2005). A l'opposé, sur la base de l'annotation de 19 400 extrémités de BAC du chromosome 3B, il a été estimé que les séquences codantes représentaient 1,2% du génome, et que le génome B portait 36 000 gènes (Paux et al., 2008). L'échantillonnage et l'annotation semblent donc avoir influencé considérablement les estimations.

L'annotation des séquences de 109 BAC par K. Devos a dressé une estimation du nombre de gènes entre 195 000 et 288 000 pour l'ensemble du génome de blé (http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0501814). Toutefois, les critères de séparation gènes/pseudogènes n'ayant pas été appliqués, cela pourrait expliquer le nombre élevé de gènes retrouvé pour chaque génome diploïde (55 000 à 111 000). Plus récemment, le séquençage de 13 régions de grandes tailles (>1 Mb en moyenne) réparties le long du chromosome 3B a permis de mettre en évidence 148 gènes et 51 pseudogènes, représentant une densité de 104 gènes par kb. En se basant sur cette densité, et en corrigeant les biais identifiés, le nombre de gènes du chromosome 3B a été estimé à 8 400, soit un total de 50 000 gènes par génome diploïde (Choulet et al. 2010). En 2011, Massa et ses collaborateurs publient une étude portant sur la dynamique de l'espace génique pendant l'évolution des génomes de *Ae. tauschii*, *B. distachyon*, *O. sativa* et *S. bicolor*. Pour cela, ils ont séquencés 9 régions du génome de *Ae. tauschii*, représentant 9,7 Mb. Grâce à ces

données, ils ont estimé le nombre total de gènes à 36 371, par extrapolation des 90 gènes annotés dans les régions séquencées (Massa et al., 2011). Les auteurs indiquent que le nombre de gènes est comparable à celui estimé en moyenne pour chaque génome à partir de séquences partielles de blé hexaploïde (données non publiées, Bennetzen JL, San Miguel P, Devos KM). Plus récemment, l'analyse de l'espace génique pour des chromosomes 3A et 5A, à partir de données de séquençage de BAC et d'extrémités de BAC, a permis d'estimer le nombre de gènes entre 5 088 et 9 571 pour les chromosomes 3A et 5A respectivement (Hernandez et al., 2012; Sehgal et al., 2012; Vitulo et al., 2011).

Pour les génomes diploïdes, l'accès à une séquence de référence a permis d'annoter les gènes codants des protéines. Pour *T. urartu*, 34 879 gènes codants des protéines ont été identifiés, d'une taille moyenne de 3 207 pb et composés de 4,7 exons en moyenne (Ling et al., 2013). La comparaison avec les valeurs estimées précédemment pour le génome A du blé tendre (E. D. Akhunov, Akhunova, & Dvořák, 2005) montre une augmentation du nombre de gènes identifiés de 6800, soit 24% de plus. Pour *Ae. tauschii*, les auteurs ont mis en évidence 34 498 gènes codant des protéines (Jia et al., 2013).

L'approche par séquençage du génome entier du blé hexaploïde a permis d'identifier le contenu en gène par similarité avec leurs orthologues (Brenchley et al., 2012). Les lectures ont été alignées sur un set de gènes de référence issu des annotations des génomes d'espèces apparentées (*B. distachyon*, *O. sativa, S. bicolor* et les ADNc de l'orge). Les lectures présentant une similarité avec le même gène de référence ont ensuite été assemblées *de novo*, afin d'obtenir les séquences des copies distinctes provenant des trois sous-génomes homéologues (méthode appelée « assemblage par groupes d'orthologues »). Une séquence ainsi assemblée est ensuite choisie comme représentante de chaque groupe d'orthologues (appelé OGR, « orthologous group representative »; (Spannagl, Martis, Pfeifer, Nussbaumer, & Mayer, 2013). Au total, 86 944 séquences orthologues ont été regroupées en 20 496 familles de gènes. Afin d'estimer le nombre de gènes, les auteurs se sont basés sur la taille moyenne des familles de gènes (s=1,46) et le taux de rétention observé (r=2,5-2,7). Les auteurs estiment ce nombre entre 93 900 et 96 300.

L'approche de l'IWGSC par tri de chromosome a permis l'annotation de 976 962 loci géniques potentiels avec 1 265 548 variants d'épissage. Au total 133 090 loci présentent une homologie avec des gènes de plantes apparentées (124 201 gènes retrouvés dans les séquences génomiques assemblées, et 8 889 transcrits non ancrés) (International Wheat Genome Sequencing Consortium, 2014). Ces loci ont été catégorisés comme étant de confiance élevée (HCS, « High Confidence Set ») et divisés en quatre classes (HCS1 à HCS4) en fonction du pourcentage de la taille du gène homologue similaire au gène prédit.

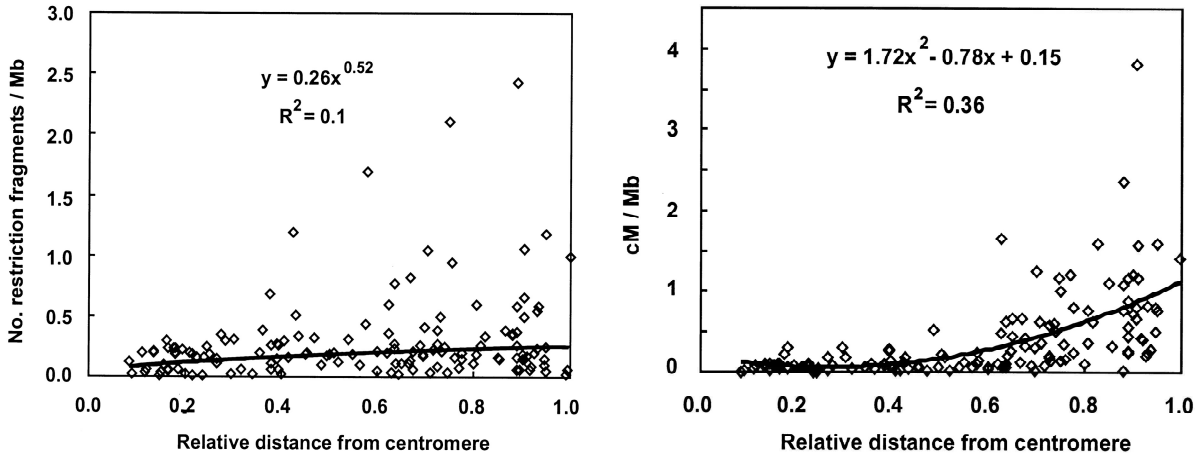



**Figure 26** : Corrélation entre la densité de gène ou le taux de recombinaison et la distance relative au centromère.
La distance relative au centromère est représentée sur l'axe x, où 0 représente de centromère et 1 les télomères. La courbe ainsi que l'équation ont été déterminées comme la meilleur approximation par rapport aux données. A gauche est représentée la densité de gène (estimée en nombre de fragments de restriction par Mb) par rapport à la distance au centromère, et à droite de taux de recombinaison (cM/Mb) par rapport au centromère. D'après (Eduard D Akhunov et al. 2003).

A l'aide d'une carte génétique à haute densité (12 966 SNPs) et en intégrant les données de synténie avec les espèces apparentées, il a été possible d'ordonner virtuellement 75 249 gènes (approche appelée « GenomeZipper » ; (Spannagl et al., 2013)). Après estimation du nombre de pseudogènes, le nombre de gènes codant une protéine a été estimé à 106 000 au sein du génome hexaploïde. A l'aide des données RNA-Seq, il a été montré que 46% des gènes HCS avaient au moins deux transcrits et en moyenne 2,6 transcrits par locus, soit environ 300 000 transcrits codant une protéine à l'échelle du génome. Les sous génomes A, B et D portent respectivement 40 253 (33%), 44 523 (35%) et 39 425 (32%) gènes. Les différences observées entre sous génomes refléteraient les différences accumulées entre ancêtres diploïdes, c'est-à-dire avant les événements de polyploïdisation.

Bien que de telles ressources soient utiles pour du développement de marqueurs SNP notamment, les informations de séquences sont très fragmentées et l'information de localisation chromosomique de tous les gènes le long des chromosomes est, soit indisponible, soit biaisée par la relation de synténie.

### 4.3.2 Organisation de l'espace génique chez le blé:

Chez le blé l'organisation des transposons en longs stretchs de séquences entraine la formation de régions riches ou pauvres en gènes, ainsi que la formation de clusters de gènes. Pour le chromosome 3B, il a été estimé que 75% des gènes sont regroupés en îlots, c'est-à-dire que ces gènes ont une distance intergénique significativement plus petite (Eduard D Akhunov et al., 2003; Choulet et al., 2010; Devos et al., 2002; Rustenholz et al., 2010). A partir de l'analyse de données de cartographie d'EST dans les bins de délétion sur les 21 chromosomes de blé tendre, une corrélation entre la densité de gènes et le taux de recombinaison le long des chromosomes a été mise en évidence (Eduard D Akhunov et al., 2003). Les régions proximales ont un faible taux de recombinaison, sont enrichies en éléments anciens, ont une faible densité de gènes, et contrastent avec les régions distales, qui ont un fort taux de recombinaison, sont enrichies en éléments plus jeunes et où la densité en gène est plus forte (Eduard D Akhunov et al., 2003). Ces travaux montrent aussi que la densité de gènes par bin augmente avec la distance par rapport au centromère (Figure 26). La notion de régions riches en gènes chez le blé avait déjà été mise en évidence (K. S. Gill, Gill, & Endo, 1993). Dans son étude, Akhunov (2003) montre que pour 21% des EST analysés, il y a un locus dupliqué sur un autre chromosome, et ces duplications sont accumulées dans les régions distales des chromosomes. Il montre aussi que la « chance » de survie du locus dupliqué portant un polymorphisme est différente dans les régions à faible ou fort taux de recombinaison. En effet, les loci issus de duplication vont être plus facilement

éliminés dans les régions à faible taux de polymorphisme, ce qui peut expliquer leur accumulation dans les régions à fort taux de polymorphisme, et par conséquent dans les parties distales des chromosomes. L'analyse des cartes physiques et des séquences des extrémités de BAC des chromosomes 1AS, 1BS, 1BL a confirmé l'existence de la distribution non uniforme des gènes. (Breen et al., 2013; Lucas et al., 2013; Paux et al., 2008). En plus de ces trois bras de chromosomes, des données de séquençage ont été publiées pour les chromosomes 4A (856 Mb), 5A (857,8 Mb), et pour le bras 3AS (Hernandez et al., 2012, Vitulo et al., 2011, Sehgal et al. 2012). La densité a été estimée, allant de 1 gène pour 105 à 162 kb pour les chromosomes 1B et 5A. Enfin pour le chromosome 1BS, les auteurs ont montré qu'il y a un gradient de densité d'un facteur deux sur l'axe centromère-télomère.

## 5 Les objectifs de la thèse

Le chromosome 3B est le plus grand des 21 chromosomes du blé tendre, avec une taille estimée en cytologie à 995 Mb. En 2004, ce chromosome fut le premier chromosome pour lequel une banque BAC a été construite (Safar et al. Plant J 2004), et, en 2008, pour lequel une carte physique a été établie (Paux et al., 2008) puis finalement optimisée (Rustenholz 2011). Cette carte couvre 97% (961 Mb) du chromosome et est composée de 1 669 contigs de BAC. Au sein du laboratoire, une puce à ADN Nimblegen a été mise au point à partir de 40 349 séquences unigènes de blé tendre présentes dans les bases de données, avec pour but d'étudier l'organisation de l'espace génique du chromosome 3B par l'hybridation des BAC composant le MTP (Rustenholz 2010 et 2011). Par cette approche, 2924 loci exprimés ont pu être cartographiés sur le chromosome 3B, confirmant l'existence d'un gradient de densité des gènes croissant le long de l'axe centromère-télomère. Cette analyse a également suggéré que 70% des gènes sont organisés en îlots (plusieurs gènes présents sur le même BAC), et que les gènes du même îlot tendent à partager un profil d'expression et/ou une fonction similaire.

Mon projet de thèse s'inscrit dans la continuité de ces travaux, ainsi que dans le cadre du projet de séquençage du chromosome 3B (ANR-France Agrimer 3BSEQ). Ce projet avait pour finalité de séquencer les 8 452 BAC couvrant ce chromosome afin de produire la première séquence de référence d'un chromosome de blé, de caractériser son espace génique en exploitant des données de séquençage RNA-Seq, et en étudiant sa composition, son organisation, son évolution à une résolution jamais atteinte auparavant pour un génome complexe. Ce projet a abouti à la construction d'une séquence unique de 774 Mb (appelée « pseudomolécule »). La pseudomolécule est composée de 1358 scaffolds ordonnés sur la base d'une carte génétique et représentant 93% des séquences assemblées pour chromosome. Ce projet est le fer de lance du Consortium IWGSC et a ouvert la voie à

l'obtention d'une séquence de référence du génome hexaploïde complet. Mon sujet de thèse s'inscrit donc dans ce contexte d'émergence de nouvelles ressources génomiques, qui constituent des avancées majeures permettant de lever les verrous qui, jusque-là, ont limité notre capacité à étudier finement l'organisation et l'expression de l'espace génique chez le blé polyploïde. Les objectifs de ma thèse sont les suivants :

(1) La construction d'une carte transcriptionnelle du chromosome 3B.
(2) La caractérisation de l'espace génique et de son expression à l'échelle du génome entier grâce aux assemblages des 21 chromosomes.

Les questions scientifiques posées concernent :

(1) Les relations entre la structure du génome, l'organisation et l'expression des gènes.
(2) La caractérisation des régions transcrites potentiellement non-codantes (« novel transcribed regions »).

# RESULTATS

Dans un premier temps, le travail de ma thèse a consisté à étudier l'espace génique dans sa globalité à l'échelle du chromosome 3B sur la base de la pseudomolécule. Pour cela, j'ai construit des librairies RNA-Seq pour 15 conditions de développement du blé. Les transcrits issus d'une première analyse ont été utilisés comme évidences biologiques dans le pipeline pour l'annotation des gènes de la pseudomolécule. Nous avons montré que les résultats obtenu sur l'organisation structurale et fonctionnelle des gènes, sont corrélés avec le profil de recombinaison. Ces résultats ont aussi été couplés à ceux obtenus avec l'analyse de l'évolution des gènes ainsi que l'annotation les éléments transposables et ont été publiés dans la revue Science (Choulet et al., 2014).

WHEAT GENOME

# Structural and functional partitioning of bread wheat chromosome 3B

Frédéric Choulet,[1,2]* Adriana Alberti,[3] Sébastien Theil,[1,2] Natasha Glover,[1,2] Valérie Barbe,[3] Josquin Daron,[1,2] Lise Pingault,[1,2] Pierre Sourdille,[1,2] Arnaud Couloux,[3] Etienne Paux,[1,2] Philippe Leroy,[1,2] Sophie Mangenot,[3] Nicolas Guilhot,[1,2] Jacques Le Gouis,[1,2] Francois Balfourier,[1,2] Michael Alaux,[4] Véronique Jamilloux,[4] Julie Poulain,[3] Céline Durand,[3] Arnaud Bellec,[5] Christine Gaspin,[6] Jan Safar,[7] Jaroslav Dolezel,[7] Jane Rogers,[8] Klaas Vandepoele,[9] Jean-Marc Aury,[3] Klaus Mayer,[10] Hélène Berges,[5] Hadi Quesneville,[4] Patrick Wincker,[3,11,12] Catherine Feuillet[1,2]

We produced a reference sequence of the 1-gigabase chromosome 3B of hexaploid bread wheat. By sequencing 8452 bacterial artificial chromosomes in pools, we assembled a sequence of 774 megabases carrying 5326 protein-coding genes, 1938 pseudogenes, and 85% of transposable elements. The distribution of structural and functional features along the chromosome revealed partitioning correlated with meiotic recombination. Comparative analyses indicated high wheat-specific inter- and intrachromosomal gene duplication activities that are potential sources of variability for adaption. In addition to providing a better understanding of the organization, function, and evolution of a large and polyploid genome, the availability of a high-quality sequence anchored to genetic maps will accelerate the identification of genes underlying important agronomic traits.

Bread wheat (Triticum aestivum L.) is a staple food for 30% of the world population. It is a hexaploid species (6x = 2n = 42, AABBDD) that originates from two interspecific hybridizations estimated to have taken place ~0.5 million and 10,000 years ago (1). The predicted closest extant representatives of the ancestral parental diploid species (2n = 14) are Triticum urartu (A genome), Aegilops speltoides (S genome related to the B genome), and Aegilops tauschii (D genome). Each of the three ancestral genomes is about 5.5 Gb in size and, therefore, results in a highly redundant 17-Gb hexaploid genome with three homeologous sets of seven chromosomes (1A to 7A, 1B to 7B, and 1D to 7D), each carrying highly similar gene copies. Moreover, most of the genome was shaped by the amplification of transposable elements (TEs) that include highly repeated families and sequences (2). This high redundancy has complicated the assembly of a complete and properly ordered reference sequence of the bread wheat genome. A fully sequenced genome enables scientists and breeders to have access to a complete gene set, with the gene order along each chromosome, and to identify candidate genes between markers associated with important traits. It also enables the identification of recent duplicates, which may be involved in species-specific evolution (3), and tracing of their evolutionary history. Before obtaining a full genome sequence, the wheat gene space has been investigated through various genome and transcriptome survey sequencing approaches and through microarray hybridizations (4–7). Recently, whole-genome shotgun sequencing of cultivar Chinese Spring using Roche/454 technology and synteny-driven assembly yielded ~95,000 gene models [N50 = 0.9 kb (8)]. Furthermore, the gene space of the diploid wild relatives Ae. tauschii (DD) and T. urartu (AA) has also been assembled and led to describe 43,150 and 34,879 genes, respectively (9, 10). Although these sequences are useful templates for marker design and comparative analyses, as a result of assembly limitations of short-read-based sequencing (11, 12), they are still very fragmented, and a large fraction of the genes are unanchored to chromosomes. The maize (Zea mays) and potato (Solanum tuberosum) sequencing projects, both representing species with highly repetitive genomes, were able to avoid overfragmentation by combining multiple sequencing technologies and through the use of DNA libraries with a diversity of insert sizes (13, 14).

The International Wheat Genome Sequencing Consortium (IWGSC) road map focuses on physically mapping and obtaining a high-quality reference sequence of each of the 21 individual wheat chromosomes rather than approaching the hexaploid genome as a whole. This strategy relies on flow-sorting individual chromosomes and/or chromosome arms from ditelosomic lines of the cultivar Chinese Spring to construct bacterial artificial chromosome (BAC) libraries (15). The largest chromosome is 3B (~1 Gb). It was the first chromosome for which a BAC library was constructed (16) and a physical map achieved (17). A pilot sequencing study on 13 contigs (2) suggested that genes tend to be mainly clustered into small islands, the presence of a twofold gene density increase from the centromere toward the telomeres, and a high proportion of nonsyntenic genes interspersed within a conserved ancestral grass gene backbone. It provided a proof of principle for this strategy and opened the way for producing a reference sequence of the large and polyploidy wheat genome.

## Sequencing and construction of a pseudomolecule

We used a hybrid sequencing and BAC pooling strategy to sequence 8452 BAC clones from the minimal tiling path (MTP) that was established during the construction of the chromosome 3B physical map (4, 18). After the integration of BAC-end sequences, manual curation of the scaffolding, gap filling, and correction of potential sequencing errors (18), we obtained a final assembly of 2808 scaffolds representing 833 Mb with a N50 of 892 kb (i.e., half of the chromosome sequence is assembled in scaffolds larger than 892 kb). We estimated that about 6% of the chromosome sequence was not present in the MTP BAC-based assembly through comparison with the 546,922 contigs assembled from whole-chromosome shotgun sequencing of flow-sorted 3B DNA (19). This suggests that the size of chromosome 3B is nearly 886 Mb—that is, about 11% smaller than originally predicted (16, 20). We built a pseudomolecule of chromosome 3B by ordering 1358 scaffolds along the chromosome using an ordered set of 2594 anchor single-nucleotide polymorphism (SNP) markers. The pseudomolecule represents 774.4 Mb (93% of the complete sequence), with a scaffold N50 of 949 kb (table S1). The order of markers was determined by linkage analysis of a recombinant inbred line (RIL) population derived from a cross between T. aestivum cultivars Chinese Spring (reference sequence) and Renan (a French elite cultivar) and refined by integrating linkage disequilibrium data from two panels and physical BAC contig information (18). This sequence corresponds to an annotation-directed improved high-quality draft (21) situated between the high-quality finished rice

[1]Institut National de la Recherche Agronomique (INRA) UMR1095, Genetics, Diversity and Ecophysiology of Cereals, 5 Chemin de Beaulieu, 63039 Clermont-Ferrand, France. [2]University Blaise Pascal, UMR1095, Genetics, Diversity and Ecophysiology of Cereals, 5 Chemin de Beaulieu, 63039 Clermont-Ferrand, France. [3]Commissariat à l'Energie Atomique et aux Énergies Alternatives, Direction des Sciences du Vivant, Institut de Génomique, Genoscope, 2 Rue Gaston Crémieux, 91000 Evry, France. [4]INRA, UR1164 Unité de Recherche Génomique Info Research Unit in Genomics-Info, INRA de Versailles, Route de Saint-Cyr, 78026 Versailles, France. [5]Centre National des Ressources Génomiques Végétales, INRA UPR 1258, 24 Chemin de Borde Rouge, 31326 Castanet-Tolosan, France. [6]Biométrie et Intelligence Artificielle, INRA, Chemin de Borde Rouge, BP 27, 31326 Castanet-Tolosan, France. [7]Centre of the Region Haná for Biotechnological and Agricultural Research, Institute of Experimental Botany, Slechtitelu 31, CZ-78371 Olomouc, Czech Republic. [8]The Genome Analysis Centre, Norwich, Norwich Research Park, Norwich NR4 7UH, UK. [9]Department of Plant Systems Biology (VIB) and Department of Plant Biotechnology and Bioinformatics (Ghent University), Technologiepark 927, 9052 Gent, Belgium. [10]Munich Information Center for Protein Sequences, Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum Muenchen, D-85764 Neuherberg, Germany. [11]CNRS UMR 8030, 2 Rue Gaston Crémieux, 91000 Evry, France. [12]Université d'Evry, CP5706 Evry, France.
*Corresponding author. E-mail: frederic.choulet@clermont.inra.fr

genome sequence and the improved draft maize genome (13).

## Annotation of genes, transcribed loci, and transposable elements

Gene modeling led to the prediction of 7264 coding loci on the 3B pseudomolecule (Table 1), including 5326 with a functional structure and 1938 (27%) likely corresponding to pseudogenes. An additional 251 gene models and 188 pseudogenes were annotated in unanchored scaffolds. RNA-Seq data revealed that 71.4% of the predicted genes/pseudogenes are transcribed and led to the identification of 3692 unannotated transcribed loci that may encode functional noncoding RNAs or unknown proteins, hereafter referred to as novel transcribed regions (NTRs) (Table 1). In addition, 791 highly conserved noncoding RNA genes involved in RNA maturation and protein synthesis were also predicted (Table 1). Chromosome 3B appears to contain a high number of small nuclear RNA genes (U1 to U6) including nine U1-snRNAs (small nuclear RNAs), seven of which are tandemly duplicated. As a comparison, there are 14 U1-snRNAs in the entire Arabidopsis thaliana genome (www.plantgdb.org). The higher number of U1-snRNAs may reflect a higher level of duplication in the wheat genome. We found 53,288 complete and 181,058 truncated copies of TEs, belonging to 485 TE families and representing 85% (640 Mb) of the 3B pseudomolecule, through a similarity-search approach. Further de novo repeat detection (18) identified 3.6% putatively new TEs.

We estimated the putative location of the centromeric region by plotting the density of the long terminal repeat retrotransposons (LTR-RTs) CRW (centromeric retrotransposons of wheat) and Quinta along the pseudomolecule. These LTR-RTs are recognized by the centromere-specific histone CenH3 and thus are centromere-functional sequences (22). Two major peaks covering a region of 122 Mb (265 to 387 Mb) (fig. S1)—which includes 1 Mb previously shown as interacting with histone CenH3 (22) and encompassing the centromere of the orthologous rice chromosome 1 (23)—were identified. This region was defined as the centromeric-pericentromeric region of chromosome 3B. A strong correlation has been observed between the size of the centromeres and the chromosomes in grasses (24), and it is likely that large chromosomes have centromeres larger than 10 Mb. This may be critical to ensure the structural rigidity of the pericentromeric regions needed for kinetochore co-orientation (25). Marker assignation to either the short or the long arm indicated the presence of a break point between 349.4 and 350.0 Mb that might be the position of the core centromere.

## Variability in recombination rate and gene density along the chromosome

We found 787 crossover (CO) events on chromosome 3B in the Chinese Spring × Renan population, with on average 2.6 COs per chromosome per individual, which is similar to maize [2.7 to 3.4 (26)]. Distribution of meiotic recombination rates revealed extreme variations along the chromosome. Whereas the average recombination rate is 0.16 cM/Mb, actual values range from 0 to 2.30 cM/Mb (per 10-Mb window) (Fig. 1A). Segmentation analysis (18) revealed partitioning with the two distal regions of 68 Mb (region R1) and 59 Mb (region R3) on the short and long arms, respectively, showing recombination rates of 0.60 cM/Mb and 0.96 cM/Mb on average, and a large proximal region of 648 Mb (region R2) spanning the centromere with an average recombination rate of 0.05 cM/Mb (Table 2 and Fig. 1A). This provides insight into the actual physical size of the highly recombinogenic regions previously detected at the end of the wheat chromosomes (27, 28). When a narrower window of 1 Mb was used, variations ranged from 0 to 12 cM/Mb (Fig. 1A), a range similar to that observed in maize [0.8 to 11.5 cM/Mb (26)] and sorghum [0 to 10 cM/Mb (29)]. All crossover events occurred in only 13% of the chromosomes in our population of 305 individuals. The largest region totally deprived of recombination corresponds to 150 Mb and includes the putative 122-Mb centromeric-pericentromeric region. This was

Table 1. General features of the 3B pseudomolecule. LINEs, long interspersed nuclear elements; SINEs, short interspersed nuclear elements; rRNA, ribosomal RNA; snoRNA, small nucleolar RNA; TIRs, terminal inverted repeats.

| Pseudomolecule sequence | | | | |
|---|---|---|---|---|
| | Length (bps) | 774,434,471 | | |
| | G+C content | 46.16% | | |
| Protein-coding genes | | All | Full genes | Pseudogenes |
| | No. of genes | 7264 | 5326 | 1938 |
| | Average size (bps) of coding sequences (T standard deviation) | 1095 T 807 | 1187 T 821 | 840 T 710 |
| | Average number of exons (T standard deviation) | 4.2 T 4.4 | 4.4 T 4.6 | 3.6 T 3.8 |
| | Gene density (kb$^{-1}$) | 10.7 | 14.5 | 40.0 |
| | No. of expressed genes | 5185 | 4125 | 1060 |
| | No. of genes with alternative splicing | 3185 | 2596 | 589 |
| | % genes with alternative splicing | 61 | 63 | 56 |
| | Average no. isoforms/expressed gene | 5.8 | 5.8 | 5.8 |
| | NTRs | 3692 | | |
| Noncoding RNA genes | | | | |
| | tRNA | 589 | | |
| | 5S rRNA | 85 | | |
| | Others (snRNA, snoRNA) | 117 | | |
| | Total | 791 | | |
| Transposable elements (TEs) | | | | |
| Class I | Copia | 15.6% | | |
| | Gypsy | 46.9% | | |
| | Unclassified LTR-retrotransposons | 3.5% | | |
| | LINEs | 1.2% | | |
| | SINEs | 0.01% | | |
| | Total class I | 67.1% | | |
| Class II | CACTA | 16.4% | | |
| | Harbinger | 0.19% | | |
| | Mariner | 0.19% | | |
| | Mutator | 0.43% | | |
| | hAT | 0.02% | | |
| | Unclassified class II with TIRs | 0.22% | | |
| | Unclassified class II | 0.10% | | |
| | Helitron | 0.01% | | |
| | Total class II | 17.6% | | |
| Unclassified repeats | | 0.81% | | |
| Total TEs | | 85.5% | | |

confirmed by the linkage disequilibrium (LD) pattern (fig. S2). Twenty-two regions showed a recombination ratio higher than 1.6 cM/Mb—i.e., >10 times the average for this chromosome—and thus may contain recombination hot spots (Fig. 1A). However, no significant correlation was observed between the recombination rate and gene content, coding DNA, or TE content of these regions.

The 7264 genes are not evenly distributed, and gene density is increasing on both arms along the centromere-telomere axis, correlating with the distance to the centromere ($r_s = 0.79$, $P < 2.2 \times 10^{-16}$, $R^2 = 0.61$) (Fig. 1B). Using a 10-Mb window, the average gene density estimate is 9 T 5 genes/Mb, ranging from 1.3 in the centromeric-pericentromeric region up to 27.9 at the most telomeric end of the short arm, a pattern commonly observed in grass genomes. Variation of the gene density in wheat chromosome 3B is higher than for chromosomes in the more compact rice genome (30); lower than in sorghum, where genes are mostly found in the telomeric regions (31); and in the same range as in maize, which also contains a high percentage of TEs (13). Segmentation analysis revealed five major regions with contrasted gene densities (Fig. 1B) and a fourfold gradient of the gene density—i.e., twice as many as suggested by the pilot study on chromosome 3B (2). The distal segments exhibiting the highest gene density (19 genes per Mb) correspond nearly to the highly recombinogenic R1 and R3 regions (Fig. 1A). The R2 region was subdivided into three segments, with the lowest gene density (5 genes per Mb) in a 234-Mb segment encompassing the centromeric-pericentromeric region. As previously suggested (2, 4), there is no large region completely devoid of coding sequence (maximum of 3.7 Mb). We confirmed that the intergenic distances (IGDs) are extremely variable (average 104 T 190 kb) and that a majority (73%) of the genes are organized in small islands, or insulae (32). This suggests that most of the intergenic regions are under selective constraint prevented from TE insertion. Indeed, only 29% of the IGDs are larger than 104 kb, but they account for 81% of the chromosome size, demonstrating that TE-mediated genome expansion likely occurred within a limited number of intergenic regions.

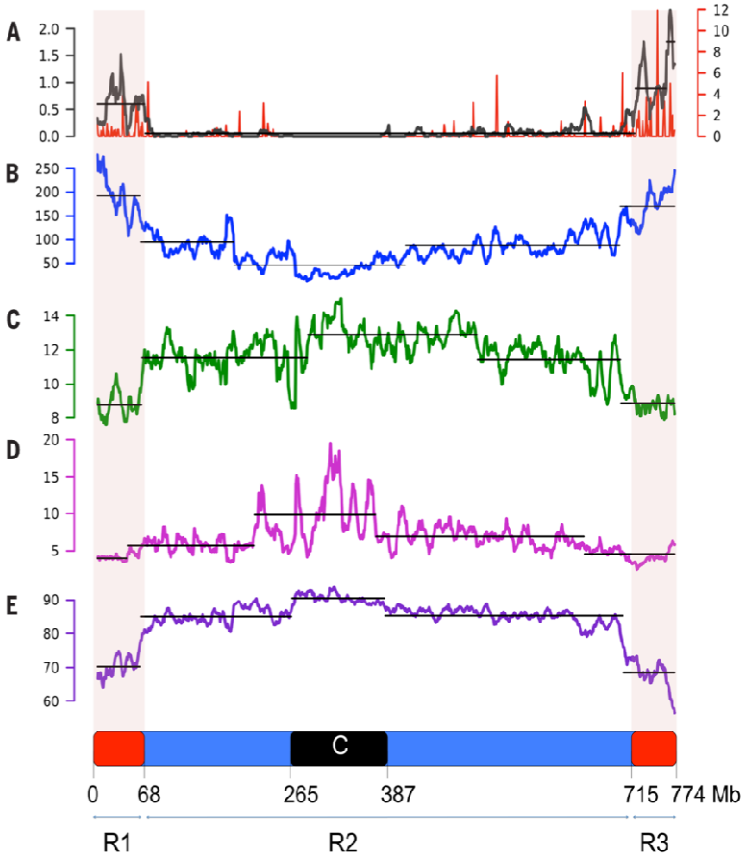



Fig. 1. Structural and functional partitioning of wheat chromosome 3B. Distribution and segmentation analysis of (A) meiotic recombination rate (cM/Mb in sliding window of 10 Mb in black and 1 Mb in red); (B) gene density (CDS/10 Mb); (C) expression breadth; (D) average number of alternative spliced transcripts per expressed gene; and (E) TE content along the 3B pseudomolecule. Distal regions of the chromosome R1 and R3 are represented in red. In (C), the centromeric/pericentromeric region is in black. The borders of these regions are indicated in Mb. Sliding window size: 10 Mb; step: 1 Mb.

## Relationships between gene expression, function, and chromosomc location

Of all annotated genes on chromosome 3B, 71.4% are expressed in at least one of the 15 conditions analyzed [five organs at three developmental stages each (table S2)], 33% in all conditions, and 5% in one only (fig. S3). On average, genes are expressed in 10.8 of 15 conditions (considering all predictions), and expressed genes are transcribed into 5.8 alternative transcripts, or isoforms. Both the expression breadth and the average number of isoforms are distributed unevenly along the chromosome, with a clear decrease of the two parameters toward the telomeres (Fig. 1, C and D). Segmentation revealed distal segments with boundaries similar to that of regions R1 and R3 and with genes expressed in fewer conditions than in the proximal region: 8.7 versus 11.7, respectively ($P < 2.2 \times 10^{-16}$, Welch t test) (Table 2). Similarly, the average number of alternative transcripts is higher in the proximal (6.5) than in the distal (4.3) regions ($P < 2.2 \times 10^{-16}$, Welch t test) (Table 2).

Gene ontology (GO) term enrichment was estimated for the R1, R2, and R3 regions (18) (tables S3 to S5). The distal regions are enriched in many GO categories, some being related to adaptation (response to abiotic stimulus or response to stress). Well-known examples of genes related to adaptation are those involved in resistance to pathogens. Chromosome 3B carries 171 genes putatively associated with disease resistance (18), and their distribution is highly biased, with 135 (79%) of them located in the distal regions (whereas these regions contain just 33% of the gene set). Such uneven distribution and the correlation with the distribution of crossovers suggest that meiotic recombination acts as a main driver for creating variability in distal regions of chromosome 3B.

To investigate whether such partitioning is a common pattern of large plant genomes, we analyzed the distribution of the gene expression breadth in maize and barley, which both exhibit large genome size (>1 Gb) and increased recombination rates at chromosomal extremities (33, 34). In barley, segmentation analysis of the seven chromosomes based on recombination data identified the same pattern as on chromosome 3B, with two highly recombinogenic distal regions and a large nonrecombinogenic region. Using expression data of eight conditions (34), we also observed that the two high-recombination distal regions carry genes expressed in fewer conditions than those carried by the low-recombination proximal regions (5.9 versus 6.7; $P = 2.2 \times 10^{-16}$, Welch t test) (fig. S4A). Using GO terms, we found a significant enrichment of these regions in the categories "cell death" and "defense response," which support previous findings that barley disease resistance genes are clustered in the distal regions (34). In contrast, in maize, although we

observed partitioning of the recombination rate, no gene-expression partitioning was detected using RNA-Seq data of 18 conditions (35). Overall, the expression breadth is 13.2 and 12.7 in high- and low-recombination regions, respectively, with chromosome specific patterns (fig. S48). Nevertheless, high-recombination regions are also enriched in GO categories "cell death" and "defense response to fungus and bacteria," suggesting that such genes are consistently found in distal recombinogenic regions in large plant genomes. These results suggest that the partitioning observed on wheat chromosome 3B is conserved in the Triticeae and may not reflect a general pattern of large genomes. Alternatively, it is possible that the active rearrangements observed in the maize genome have modified this pattern. Additional evidence will come once other large plant genomes (>1 Gb) are sequenced and analyzed.

### Uneven distribution of transposable elements

LTR-RTs represent 66% of the chromosome 3B sequence (gypsy, 47%; copia, 16%; unclassified, 3%) (Table 1), which is slightly lower than the ~75% of LTR-RT identified in the whole maize genome (36). Only 4% (3 out of 85) of the LTR-RT families were found in single copies, compared with 41% in maize (36) and 48% in the rice genome (37). Sixteen percent of the sequence is composed of class II DNA transposons that mostly correspond to CACTA elements (Table 1), compared with 3.2% in the maize genome (13). Only six families account for 50% of the wheat chromosome 3B TE fraction, as previously suggested from partial sequence analyses (2, 38) and from observations in other large genomes (36). However, in contrast to the maize genome, in which most of the intact elements are found in 1 to 10 copies (36), the majority of the TE families annotated on chromosome 3B have a higher number of copies (10 to 1000 copies). Estimated insertion dates for the most abundant LTR-RT families showed a major peak at 1.5 million years (My) but also quite specific patterns of TE activity for each family (fig. S5). Our data support the hypothesis (2, 38) that most of the transposable elements that shaped the B genome were inserted before polyploidization [0.5 million years ago (Ma)] and have been less active since then. Distribution of recently inserted elements revealed that TE insertion occurred at a similar rate in the distal and proximal regions. In contrast, older insertions (>1.5 Ma) were 1.7 times as abundant in the R2 region compared with the R1 and R3 regions, suggesting a higher rate of TE elimination in the distal ends of chromosome 3B.

The TE density distribution was not random (Fig. 1E), with a lower density in the R1 (73%) and R3 (68%) regions compared with the R2 region (88%) (Table 2). The 122-Mb centromeric-pericentromeric region displayed the highest density (93%) of TEs. Beneath the global TE distribution pattern, each superfamily presents its own specificities (fig. S6). For example, CACTA transposons are more abundant in the distal gene-rich regions (Table 2), supporting in situ hybridization findings at the whole-genome level (39). In addition, the distribution of TE families varied on the basis of their relative distance to genes (18) (fig. S7). DNA transposons Mutator, Harbinger, and MITEs are found close to genes, whereas LTR RTs and CACTAs tend to be located at much larger distances from the genes. For instance, the 17,479 annotated MITEs were found to be significantly associated with genes ($r = 0.89$; $P < 1 \times 10^{-10}$), as previously observed in plant genomes (40).

### Synteny between chromosome 3B and related grass genomes

Comparative genomics in grasses has been used to define syntenic relationships between different species (41, 42) and to provide insight into their evolution since the divergence from a common ancestor 50 to 70 Ma (43). We compared the wheat chromosome 3B genes (Ta3B) with the closest sequenced relative, Brachypodium distachyon [common ancestor, 32 to 39 Ma (44)], and with one representative of each of the Ehrhartoideae and Panicoideae grass subfamilies: Oryza sativa indica [rice (30)] and Sorghum bicolor (31), respectively. Wheat chromosomes of group 3 are syntenic with chromosome 1 of rice (Os1), chromosome 3 of sorghum (Sb3), and the distal parts of B. distachyon chromosome 2 (Bd2). We first investigated potential gene loss after polyploidization by using the conserved and syntenic genes found on chromosomes Os1, Bd2, and Sb3. These represent the grass core genes that are expected to be present on wheat homoeologous group 3, unless they have been lost by fractionation after polyploidization. The finding that 94% of the conserved genes are also present on the 3B sequence (Fig. 2A), which represents 94% of the chromosome (see above), suggests that no major gene loss has occurred in the B subgenome yet. This is confirmed at the whole-genome level by the results of the chromosome survey sequences (19). In contrast, 2065 genes on chromosome 3B (34.6%, including pseudogenes) shared similarity with genes on nonorthologous chromosomes in the other grass genomes. This proportion of nonsyntenic genes is much higher than the 5% (between 149 and 207) of nonsyntenic genes found in the other grass species analyzed (Fig. 2A and table S6). It confirms previous results showing substantial modifications and rearrangements of the wheat gene space (2). When looking at the conservation of the gene order, collinear genes represent 42 to 68% of the genes present on Os1, Bd2, and Sb3, whereas they represent less than 30% of the Ta3B genes (including pseudogenes) (table S7 and fig. S8). The spatial distribution of syntenic and nonsyntenic genes

Table 2. Distribution of features in the three regions of chromosome 3B as defined from the recombination segmentation along the chromosome.

| | R1 | R2 | R3 |
|---|---|---|---|
| Size (Mb) | 68 | 648 | 59 |
| Recombination rate (cM/Mb) | 0.60 | 0.05 | 0.96 |
| Genes | | | |
| Predicted gene density ($Mb^{-1}$) | 19 | 7 | 19 |
| Number of predicted genes/pseudogenes | 1318 | 4845 | 1,101 |
| Full genes (no.) | 910 (69%) | 3682 (76%) | 734 (67%) |
| Pseudogenes/gene fragments (no.) | 408 (31%) | 1163 (24%) | 367 (33%) |
| Mean intergenic distance (kb) | 49 | 130 | 52 |
| Expressed predicted genes (no.) | 823 (62%) | 3629 (75%) | 733 (67%) |
| Expressed full genes (no.) | 621 | 2963 | 541 |
| Expressed pseudogenes/fragments (no.) | 202 | 666 | 192 |
| Average expression breadth (per expressed gene; /15) | 8.8 | 11.7 | 8.6 |
| Average FPKM (per expressed gene) | 141 | 255 | 156 |
| Average number of isoforms (per expressed gene) | 4.2 | 6.5 | 4.4 |
| Proportion of nonsyntenic genes* (%) | 44 | 28 | 53 |
| Proportion of intrachromosomally duplicated genes* (%) | 49 | 33 | 42 |
| Proportion of tandemly duplicated genes* (%) | 24 | 14 | 22 |
| Proportion of dispersed duplicated genes* (%) | 26 | 18 | 20 |
| Proportion of interchromosomally duplicated genes* (%) | 36 | 33 | 37 |
| Transposable elements (%) | 73.0 | 88.3 | 68.4 |
| Copia (%) | 14.7 | 15.8 | 14.1 |
| Gypsy (%) | 31.7 | 50.3 | 27.1 |
| CACTA (%) | 18.7 | 15.9 | 19.5 |

*Number of duplicated genes (filtered set, including pseudogenes) divided by the total number of genes in each region.

along the 3B pseudomolecule (Fig. 2B) shows an increased proportion of nonsyntenic genes in the R1 (44%) and R3 (53%) regions compared with the R2 region (28%) (Table 2). This supports the hypothesis that accelerated evolution occurred in the wheat lineage compared with other grasses (2, 45, 46), with insertions of nonsyntenic genes intercalated in the ancestral grass genome backbone through gene duplications or translocations that preferentially occurred in the distal recombining regions.

## Origin and evolution of nonsyntenic genes

With such a high proportion of nonsyntenic genes, one key question is whether these genes are under selection pressure or in the process of becoming pseudogenes. On the basis of the coding sequence structure, 32% of the nonsyntenic genes (versus 17% of syntenic genes) were annotated as likely pseudogenes or gene fragments. This ratio is not surprising, given that TE activity can duplicate gene fragments that are dead upon arrival. Expression patterns revealed that a majority of the nonsyntenic genes (69% versus 82% of syntenic genes) are expressed in at least one condition tested (table S8), thereby suggesting that a large fraction of these relocated genes are unlikely to be pseudogenes and may contribute to recent wheat genome evolution and, therefore, to adaptation. Interestingly, a majority (51%) of the genes expressed in a single condition corresponds to nonsyntenic genes, whereas 80% of the genes that are expressed in all 15 conditions are syntenic genes (fig. S9). This suggests that nonsyntenic genes are involved in specific processes that may be related to adaptation, whereas syntenic genes tend to be associated with essential biological processes. This hypothesis is supported by the fact that putative resistance genes identified on chromosome 3B are mainly nonsyntenic genes (18). In addition, GO term enrichment of nonsyntenic genes revealed an overrepresentation of genes involved in response to stress (table S9).

The fact that chromosome 3B exhibits a higher number of genes than its orthologs in other grasses and that at least 94% of the ancestral grass gene backbone is conserved indicates that most insertions of nonsyntenic genes result from interchromosomal duplication with retention of the parental copy. To test this hypothesis, we used the sequences of the 18 bread wheat chromosomes nonhomeologous to group 3 chromosomes (19) to search for potential parental copies of chromosome 3B genes elsewhere in the genome (18) (table S10). A paralog was identified for 87% of the nonsyntenic genes (18), with no bias regarding the chromosomal origin of the interchromosomally duplicated genes (fig. S10). Duplications of DNA fragments to different locations in a genome have been shown to result from double-strand break (DSB) repair (in which a copy of the foreign DNA is used as filler to repair the break) or capture by active TEs (46, 47). We analyzed the composition of the regions flanking the syntenic versus nonsyntenic genes (20 kb on each side) and found a high association of nonsyntenic genes with a class II transposon superfamily: 41% more CACTAs were found around nonsyntenic genes than around syntenic genes (fig. S11). CACTA transposons are known to capture genes (31, 48) and may have contributed substantially to interchromosomal gene duplications in wheat.

We also investigated the time since duplication of nonsyntenic genes through the analysis of nucleotide substitution rates ($K_s$) (18). In total, 63% of these duplications were older than 10 My and, thus, are likely shared within other Triticeae species, whereas 37% are potentially wheat-specific. Comparison with the barley genome survey sequence data (34) showed that at least 29% of the 3B nonsyntenic genes (versus 51% of the syntenic genes) are orthologous with barley chromosome 3H, confirming that part of the nonsyntenic genes were relocated before the divergence of wheat and barley 10 to 14 Ma.

We next asked if the high gene duplication activity is also observed at the intrachromosomal level. We identified 809 gene families with two or more copies comprising 2216 genes on chromosome 3B, which is about three times as much as in rice, Brachypodium, and sorghum (table S11). This indicates that, in proportion, more than twice as many genes were duplicated or retained after intrachromosomal duplications in wheat (~37%) compared with the other three grasses (~15 to 18%). About 46% of the duplicated genes of chromosome 3B are found in tandem, whereas 54% are dispersed duplicates (18) (table S12). In other grass species, a majority of the duplicated genes are organized in tandem. Given the high interchromosomal duplication activity observed in our analyses (see above), it is possible that some dispersed duplicates on chromosome 3B originated through independent interchromosomal duplications rather than through intrachromosomal duplications, thereby leading to overestimates of the latter. However, even when considering syntenic dispersed duplicates—i.e., those genes that have remained at their ancestral locus and have undergone intrachromosomal duplication—23% of the whole gene set appears to have originated from recent intrachromosomal duplications, which is still higher than in other grass species. Thus, we conclude that both inter- and intrachromosomal

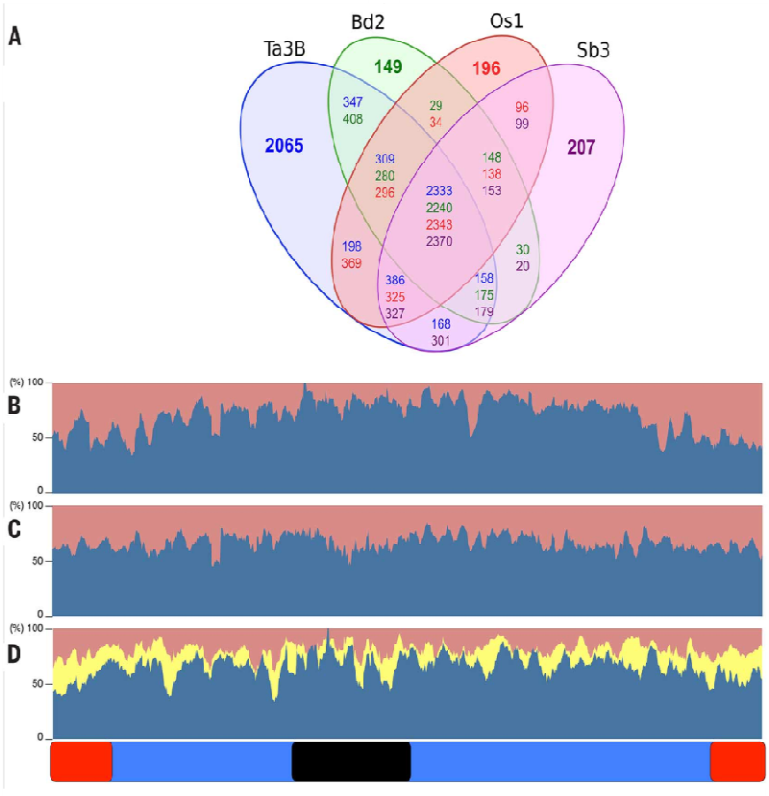



Fig. 2. Inter- and intraspecific comparative analyses of the gene content of wheat chromosome 3B. (A) Venn diagram displaying the number of genes conserved between wheat chromosome 3B (Ta3b, blue) and orthologous chromosomes in rice (Os1, red), Brachypodium (Bd2, green), and sorghum (Sb3, purple). The number of nonsyntenic genes is indicated in bold for each species. Distribution along the 774 Mb of the chromosome 3B pseudomolecule of the relative proportion of (B) syntenic (blue) versus nonsyntenic (red) genes, (C) interchromosomal duplications (duplicates in red, group 3–specific genes in blue), and (D) tandem (yellow) and dispersed (red) intrachromosomal duplications and singletons (blue). Chromosome 3B is represented at the bottom with distal regions in red and the centromeric/pericentromeric region in black.

rates of duplication are higher in wheat than in the other grass species analyzed so far. Interestingly, interchromosomal duplicates were distributed uniformly along the chromosome, whereas the proportion of tandem duplicates slightly increased in the distal regions (Fig. 2D). This suggests that long-distance and tandem duplications likely arose through different mechanisms. Finally, expression analysis of intrachromosomal duplicated genes indicated that 49% of the families show expression of all copies in at least one condition. Similar to what was observed for the interchromosomal duplicated genes, the intrachromosomal duplicated genes tend to be expressed in fewer conditions as compared with nonduplicated genes (fig. S9 and table S8), suggesting that they may be undergoing subfunctionalization.

### QTL mining

As exemplified in rice and other crops, a reference genome sequence provides a resource for gene discovery, marker development, and allele mining in support of crop improvement (49). We identified 153,190 insertion site-based polymorphism (ISBP) markers (50) and 35,579 microsatellite markers along the 3B chromosome. We also located 121 quantitative trait loci (QTLs) for 50 different traits on chromosome 3B (table S13). Using these data, we conducted a meta-analysis that integrates QTLs defined in independent studies (51) and identified 18 metaQTLs with confidence intervals covering between 1.5 and 620 Mb of the chromosome 3B sequence. The largest one encompasses the centromeric region, where recombination is suppressed. Five metaQTLs with small intervals (<10 Mb) that contain between 23 and 266 protein-coding genes and between 511 and 4049 markers are suitable for fine mapping (table S14).

## Discussion

We present a reference sequence of chromosome 3B that can be used to precisely delineate structural and functional features along a chromosome and establish correlations between recombination intensity, gene density, gene expression, and evolution rate. Our results indicate that during evolution, regions with distinct features become delineated along chromosome 3B, including relatively small distal regions that are preferential targets for recombination, adaptation, and genomic plasticity. Whether our observations reflect a general pattern for the wheat genome will need to be confirmed by the analysis of other chromosome reference sequences. Already, some of the features—such as the CACTA distribution, the high rate of intrachromosomal duplication, the absence of major gene loss since polyploidization, and the gradient of gene density—have been confirmed at the whole-genome level (19, 39). Moreover, the ordered chromosome 3B sequence allowed us to distinguish duplicated genes and provided evidence for superimposed mechanisms of gene duplications. The high level of gene duplication (allopolyploidy and inter- and intrachromosomal duplications) provides the wheat genome with a vast reservoir of functional genes that likely contribute to wheat adaptation.

On the basis of this work, the IWGSC has already defined an adapted BAC pooling strategy to reach the same sequence quality while reducing sequencing costs for the remaining chromosomes. Although progress in sequencing technologies and cost reduction allows for more cost-efficient sequence production, the challenge of bioinformatics and limitations of current sequencing technologies remain (12). Solving these issues and improving methods to efficiently anchor and orient scaffolds within pseudomolecules should make the assembly of high-quality reference sequences of complex genomes routine work in the future. There is no doubt that, as witnessed after the release of the rice genome sequence (49), the number of genes cloned from wheat will grow exponentially in the near future, thereby enabling wheat researchers and breeders to cope with the urgent need to improve wheat yield in the face of climate change and food-security challenges (52).

## Materials and methods

### Sequencing, assembly, and scaffolding

A total of 8452 BACs representing the MTP of wheat chromosome 3B were pooled into 922 BAC pools. Each pool was used to create a barcoded Roche/454 8-kb–long paired-end library. In total, 150 sequencing runs were performed, leading to an average of 36-fold sequence coverage. After assembly with Newbler (Roche), we integrated 42,551 BAC end sequences to validate and improve scaffolding. Illumina reads generated from sorted DNA of chromosome 3B were used to fill gaps within scaffolds and correct potential sequencing errors remaining in the consensus sequence (18).

### Anchoring scaffolds

SNP discovery was performed through sequence capture for 52,265 loci flanking TE junctions representing an average density of one locus per 16.2 kb (18). Out of 39,077 SNPs distributed along the chromosome, a subset of 3075 evenly distributed (38.2 T 9.4 SNPs per 10 Mb) SNPs was selected to genotype 1025 lines from recombinant inbred and association panels. An anchor genetic map was built first by linkage analysis and integration of linkage disequilibrium data. A consensus map comprising 5318 markers was also built using 40 different genetic maps to anchor additional scaffolds (18). Finally, a position in the pseudomolecule was inferred for scaffolds without marker information but belonging to an anchored physical BAC contig.

### Sequence annotation

Gene modeling was performed using an improved version of the TriAnnot pipeline (53). Noncoding RNA genes were predicted using three different programs (18), and predictions were manually curated. Predictions of TEs and reconstruction of the pattern of nested insertions were performed through the development of a specific program (18) that automatically curates similarity-search results obtained with a dedicated databank comprising 4929 known wheat TEs classified into 521 families.

### Gene expression analyses

Thirty RNA samples, corresponding to RNAs extracted in duplicates from five organs (root, leaf, stem, spike, and grain) at three developmental stages each from hexaploid wheat cultivar Chinese Spring (4), were used for gene expression analyses. RNA-Seq libraries were constructed using the IlluminaTruSeq (Illumina, CA, USA) RNA sample preparation kit and sequenced. An average of 50 T 11 million paired-end reads per sample were mapped on the chromosome 3B scaffolds and used to reconstruct transcripts and estimate transcript abundance in units of fragments per kb of exon per million mapped reads (FPKM). Regions with FPKM values higher than zero were considered as expressed.

### Distribution and segmentation analyses

Distributions of recombination rate, gene and TE densities, and expression breadth were calculated within a sliding window of 10 Mb (and 1 Mb for the recombination rate), with a step of 1 Mb along the chromosome sequence using a homemade Perl script. Segmentation analyses of these distributions were performed using the R package changepoint v1.0.6 (54), with Segment Neighborhoods method and Bayesian information criterion penalty on the mean change.

### Comparative genomics, gene duplications, and molecular evolution

We performed an all-by-all Basic Local Alignment Search Tool for Proteins (BLASTP) [cutoff expected value (Evalue), $1 \times 10^{-5}$] comparison between the amino acid sequences of predicted genes of wheat chromosome 3B, rice (Michigan State University version 7.0), Brachypodium (Brachypodium Sequencing Initiative, 2.0), and sorghum (phytozome, version 1.4). We filtered out genes with no homology with at least one other gene in a compared species (cutoff 35% amino acid identity and 35% sequence overlap). Syntenic genes were defined as genes with a best BLAST hit on an orthologous chromosome in at least one other species. Nonsyntenic genes were defined as genes for which the best BLAST hit was on a nonorthologous chromosome in the other species. Clustering of orthologous and paralogous genes was performed using OrthoMCL [Evalue cutoff, $1 \times 10^{-5}$; percentage match cutoff, 35% (55)]. All 3B genes clustered into the same family were considered intrachromosomal duplicates. 3B genes clustered in a family with wheat gene models annotated on another chromosome (19), not including genes from group 3, were considered as interchromosomal duplicates. Tandem duplicates were defined as genes in the same family with five or fewer spacer genes separating them on the pseudomolecule, and dispersed duplicates were defined as having more than five spacer genes. Synonymous ($K_s$) and nonsynonymous ($K_a$) substitution rates were calculated based on ClustalW 2.1 (56) coding sequence alignments by the Nei and Gojobori method

using codeml [part of the PAML package (57)]. Age of gene divergence was estimated by the equation $K_s/2r$, where $r = 6.5 \times 10^{-9}$.

## REFERENCES AND NOTES


1. J. Dubcovsky, J. Dvorak, Genome plasticity a key factor in the success of polyploid wheat under domestication. Science 316, 1862–1866 (2007). doi: 10.1126/science.1143986; pmid: 17600208
2. F. Choulet et al., Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. Plant Cell 22, 1686–1701 (2010). doi: 10.1105/tpc.110.074187; pmid: 20581307
3. J. Zhang, Evolution by gene duplication: An update. Trends Ecol Evol 18, 292–298 (2003). doi: 10.1016/S0169-5347(03)00033-8
4. C. Rustenholz et al., A 3,000-loci transcription map of chromosome 3B unravels the structural and functional features of gene islands in hexaploid wheat. Plant Physiol. 157, 1596–1608 (2011). doi: 10.1104/pp.111.183921; pmid: 22034626
5. L. D. Wilson et al., A transcriptomics resource for wheat functional genomics. Plant Biotechnol. J. 2, 495–506 (2004). doi: 10.1111/j.1467-7652.2004.00095.x; pmid: 17147622
6. P. R. Bhat et al., Mapping translocation breakpoints using a wheat microarray. Nucleic Acids Res. 35, 2936–2943 (2007). doi: 10.1093/nar/gkm148; pmid: 17439961
7. L. L. Qi et al., A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. Genetics 168, 701–712 (2004). doi: 10.1534/genetics.104.034868; pmid: 15514046
8. R. Brenchley et al., Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature 491, 705–710 (2012). doi: 10.1038/nature11650; pmid: 23192148
9. J. Jia et al., Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. Nature 496, 91–95 (2013). doi: 10.1038/nature12028; pmid: 23535592
10. H. Q. Ling et al., Draft genome of the wheat A-genome progenitor Triticum urartu. Nature 496, 87–90 (2013). doi: 10.1038/nature11997; pmid: 23535596
11. M. C. Schatz, A. L. Delcher, S. L. Salzberg, Assembly of large genomes using second-generation sequencing. Genome Res. 20, 1165–1173 (2010). doi: 10.1101/gr.101360.109; pmid: 20508146
12. V. Marx, Next-generation sequencing: The genome jigsaw. Nature 501, 263–268 (2013). doi: 10.1038/501263a; pmid: 24025842
13. P. S. Schnable et al., The B73 maize genome: Complexity, diversity, and dynamics. Science 326, 1112–1115 (2009). doi: 10.1126/science.1178534; pmid: 19965430
14. X. Xu et al., Genome sequence and analysis of the tuber crop potato. Nature 475, 189–195 (2011). doi: 10.1038/nature10158; pmid: 21743474
15. J. Doležel, M. Kubaláková, E. Paux, J. Bartoš, C. Feuillet, Chromosome-based genomics in the cereals. Chromosome Res. 15, 51–66 (2007). doi: 10.1007/s10577-006-1106-x; pmid: 17295126
16. J. Safár et al., Dissecting large and complex genomes: Flow sorting and BAC cloning of individual chromosomes from bread wheat. Plant J. 39, 960–968 (2004). doi: 10.1111/j.1365-313X.2004.02179.x; pmid: 15341637
17. E. Paux et al., A physical map of the 1-gigabase bread wheat chromosome 3B. Science 322, 101–104 (2008). doi: 10.1126/science.1161847; pmid: 18832645
18. Supplementary materials are available on Science Online.
19. International Wheat Genome Sequencing Consortium, A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science 345, 1251788 (2014).
20. B. S. Gill, B. Friebe, T. Endo, Standard karyotype and nomenclature system for description of chromosome bands and structural aberrations in wheat (Triticum aestivum). Genome 34, 830–839 (1991). doi: 10.1139/g91-128
21. P. S. Chain et al., Genomics. Genome project standards in a new era of sequencing. Science 326, 236–237 (2009). doi: 10.1126/science.1180614; pmid: 19815760
22. B. Li et al., Wheat centromeric retrotransposons: The new ones take a major role in centromeric structure. Plant J. 73, 952–965 (2013). doi: 10.1111/tpj.12086; pmid: 23253215
23. H. Yan et al., Intergenic locations of rice centromeric chromatin. PLOS Biol. 6, e286 (2008). doi: 10.1371/journal.pbio.0060286; pmid: 19067486
24. H. Zhang, R. K. Dawe, Total centromere size and genome size are strongly correlated in ten grass species. Chromosome Res. 20, 403–412 (2012). doi: 10.1007/s10577-012-9284-1; pmid: 22552915
25. T. Sakuno, K. Tada, Y. Watanabe, Kinetochore geometry defined by cohesion within the centromere. Nature 458, 852–858 (2009). doi: 10.1038/nature07876; pmid: 19370027
26. Q. Pan, F. Ali, X. Yang, J. Li, J. Yan, Exploring the genetic characteristics of two recombinant inbred line populations via high-density SNP markers in maize. PLOS ONE 7, e52777 (2012). doi: 10.1371/journal.pone.0052777; pmid: 23300772
27. A. J. Lukaszewski, C. A. Curtis, Physical distribution of recombination in B-genome chromosomes of tetraploid wheat. Theor. Appl. Genet. 86, 121–127 (1993). doi: 10.1007/BF00223816; pmid: 24193391
28. C. Saintenac et al., Detailed recombination studies along chromosome 3B provide new insights on crossover distribution in wheat (Triticum aestivum L.). Genetics 181, 393–403 (2009). doi: 10.1534/genetics.108.097469; pmid: 19064706
29. J. Evans et al., Extensive variation in the density and distribution of DNA polymorphism in sorghum genomes. PLOS ONE 8, e79192 (2013). doi: 10.1371/journal.pone.0079192; pmid: 24265758
30. International Rice Genome Sequencing Project, The map-based sequence of the rice genome. Nature 436, 793–800 (2005). doi: 10.1038/nature03895; pmid: 16100779
31. A. H. Paterson et al., The Sorghum bicolor genome and the diversification of grasses. Nature 457, 551–556 (2009). doi: 10.1038/nature07723; pmid: 19189423
32. A. Gottlieb et al., Insular organization of gene space in grass genomes. PLOS ONE 8, e54101 (2013). doi: 10.1371/journal.pone.0054101; pmid: 23326580
33. M. W. Ganal et al., A large maize (Zea mays L.) SNP genotyping array: Development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. PLOS ONE 6, e28334 (2011). doi: 10.1371/journal.pone.0028334; pmid: 22174790
34. K. F. Mayer et al., A physical, genetic and functional sequence assembly of the barley genome. Nature 491, 711–716 (2012). pmid: 23075845
35. R. S. Sekhon et al., Maize gene atlas developed by RNA sequencing and comparative evaluation of transcriptomes based on RNA sequencing and microarrays. PLOS ONE 8, e61005 (2013). doi: 10.1371/journal.pone.0061005; pmid: 23637782
36. R. S. Baucom et al., Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. PLOS Genet. 5, e1000732 (2009). doi: 10.1371/journal.pgen.1000732; pmid: 19936065
37. R. S. Baucom, J. C. Estill, J. Leebens-Mack, J. L. Bennetzen, Natural selection on gene function drives the evolution of LTR retrotransposon families in the rice genome. Genome Res. 19, 243–254 (2009). doi: 10.1101/gr.083360.108; pmid: 19029538
38. M. Charles et al., Dynamics and differential proliferation of transposable elements during the evolution of the B and A genomes of wheat. Genetics 180, 1071–1086 (2008). doi: 10.1534/genetics.108.092304; pmid: 18780739
39. E. M. Sergeeva, E. A. Salina, I. G. Adonina, B. Chalhoub, Evolutionary analysis of the CACTA DNA-transposon Caspar across wheat species using sequence comparison and in situ hybridization. Mol. Genet. Genomics 284, 11–23 (2010). doi: 10.1007/s00438-010-0544-5; pmid: 20512353
40. C. Lu et al., Miniature inverted-repeat transposable elements (MITEs) have been accumulated through amplification bursts and play important roles in gene expression and species diversity in Oryza sativa. Mol. Biol. Evol. 29, 1005–1017 (2012). doi: 10.1093/molbev/msr282; pmid: 22096216
41. M. D. Gale, K. M. Devos, Comparative genetics in the grasses. Proc. Natl. Acad. Sci. U.S.A. 95, 1971–1974 (1998). doi: 10.1073/pnas.95.5.1971; pmid: 9482816
42. K. M. Devos, M. D. Gale, Genome relationships: The grass model in current research. Plant Cell 12, 637–646 (2000). doi: 10.1105/tpc.12.5.637; pmid: 10810140
43. F. Murat et al., Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. Genome Res. 20, 1545–1557 (2010). doi: 10.1101/gr.109744.110; pmid: 20876790
44. International Brachypodium Initiative, Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature 463, 763–768 (2010). doi: 10.1038/nature08747; pmid: 20148030
45. E. D. Akhunov et al., Synteny perturbations between wheat homoeologous chromosomes caused by locus duplications and deletions correlate with recombination rates. Proc. Natl. Acad. Sci. U.S.A. 100, 10836–10841 (2003). doi: 10.1073/pnas.1934431100; pmid: 12960374
46. T. Wicker et al., Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives. Plant Cell 23, 1706–1718 (2011). doi: 10.1105/tpc.111.086629; pmid: 21622801
47. T. Wicker, J. P. Buchmann, B. Keller, Patching gaps in plant genomes results in gene movement and erosion of colinearity. Genome Res. 20, 1229–1237 (2010). doi: 10.1101/gr.107284.110; pmid: 20530251
48. M. Morgante et al., Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat. Genet. 37, 997–1002 (2005). doi: 10.1038/ng1615; pmid: 16056225
49. C. Feuillet, J. E. Leach, J. Rogers, P. S. Schnable, K. Eversole, Crop genome sequencing: Lessons and rationales. Trends Plant Sci. 16, 77–88 (2011). doi: 10.1016/j.tplants.2010.10.005; pmid: 21094078
50. E. Paux et al., Insertion site-based polymorphism markers open new perspectives for genome saturation and marker-assisted selection in wheat. Plant Biotechnol. J. 8, 196–210 (2010). doi: 10.1111/j.1467-7652.2009.00477.x; pmid: 20078842
51. B. Goffinet, S. Gerber, Quantitative trait loci: A meta-analysis. Genetics 155, 463–473 (2000). pmid: 10790417
52. J. A. Foley et al., Solutions for a cultivated planet. Nature 478, 337–342 (2011). doi: 10.1038/nature10452; pmid: 21993620
53. P. Leroy et al., TriAnnot: A versatile and high performance pipeline for the automated annotation of plant genomes. Front. Plant Sci. 3, 5 (2012). doi: 10.3389/fpls.2012.00005; pmid: 22645565
54. J. Chen, A. K. Gupta, Parametric Statistical Change Point Analysis: With Applications to Genetics, Medicine, and Finance (Birkhäuser, Basel, 2012).
55. L. Li, C. J. Stoeckert Jr., D. S. Roos, OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res. 13, 2178–2189 (2003). doi: 10.1101/gr.1224503; pmid: 12952885
56. J. D. Thompson, D. G. Higgins, T. J. Gibson, CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680 (1994). doi: 10.1093/nar/22.22.4673; pmid: 7984417
57. Z. Yang, PAML: A program package for phylogenetic analysis by maximum likelihood. CABIOS 13, 555–556 (1997). pmid: 9367129


## ACKNOWLEDGMENTS


The authors thank the scientific advisory board (P. Schnable, S. Rounsley, D. Ware, J. Rogers, and K. Eversole) of the 3BSEQ project for fruitful discussions; K. Eversole for critical reading and editing of the manuscript; H. Rimbert, N. Cubizolles, and E. Rey for SNP marker discovery and genotyping; M. Kubaláková and J. Vrána for assistance with the preparation of DNA amplified from flow-sorted chromosome 3B; L. Couderc, A. Kellet, and S. Reboux for their support in database and system administration; and C. Poncet and the "Plateforme GENTYANE" for SNP genotyping. This work was supported by a grant from the French National Research Agency (ANR-09-GENM-025 3BSEQ), a grant from France Agrimer, and a grant (project: DI-BLE) from the INRA HAP Biologie et Amélioration des Plantes division. M.G. is funded by a grant from the European Commission research training program Marie Curie Actions (HPAI-MC-IIF-NoncollinearGenes). J. Daron is funded by a grant from the French Ministry of Research. L.P. is funded by a grant from the Region Auvergne. K.V. is supported by the Ghent University Multidisciplinary Research Partnership ("Bioinformatics: From nucleotides to networks" (Project: OMRG3101W)). J. Doležel is supported by the Czech Science Foundation (award no. P501/12/G090). The chromosome 3B BAC library and the pools of the MTP BAC clones are available upon request under a materials transfer agreement at the French IPT Genomic Center, INRA-Centre National de Ressources Génomiques Végétales. Annotation data and browser are available at https://urgi.versailles.inra.fr/gbrowse/wheat_annot_3B. Sequences and annotations of the reference pseudomolecule and unassigned scaffolds have been deposited in ENA (project PRJEB4376) under accession numbers HG670306 and CBUC01000001 to CBUC01000450, respectively. RNA-Seq data were deposited under accession number ERP004714.


## SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/345/6194/1249721/suppl/DC1
Materials and Methods
Figs. S1 to S11
Tables S1 to S14
References (58–125)


13 December 2013; accepted 30 May 2014
10.1126/science.1249721

**Conclusions Article n°1**

Dans cet article, nous montrons qu'à partir d'une approche chromosome-spécifique, nous avons pu assembler et annoter la première séquence de référence d'un chromosome de blé hexaploïde, le chromosome 3B. Cette séquence représente 774 Mb et est composée de 85% d'éléments transposables. L'utilisation de programme TriAnnot a permis de prédire 5 326 gènes et 1 938 pseudogènes, dont le gradient de densité augmente sur l'axe centromère-télomère.

L'analyse de la distribution de la recombinaison, ainsi que de la densité de gènes et d'éléments transposables par segmentation a montré que le chromosome était scindé en trois régions : deux régions distales d'environ 60 Mb où se produit la quasi-totalité de la recombinaison et où la densité de gènes est la plus forte, et une région proximale où la recombinaison est presque absente, et la densité de gènes la plus faible.

La production et l'analyse de données RNA-Seq pour 15 conditions de développement a permis de mettre en évidence l'expression de 71,4% des prédictions dans au moins une des conditions. Ces gènes sont exprimés en moyenne dans 10,8 conditions et sous forme de 5,8 transcrits alternatifs. En couplant ces données aux régions précédemment définis, nous avons constaté que ces régions différaient également en termes d'amplitude d'expression et d'épissage alternatif des gènes qu'elles contiennent, avec notamment une plus grande spécificité des régions distales. En complément des prédictions, nous avons détecté 3 692 régions transcriptionnellement actives.

Afin d'explorer en détail l'organisation structurale et fonctionnelle de ce chromosome, nous avons exploité les données de RNA-Seq pour étudier les relations entre la structure des gènes et du génome et leur régulation.

# Deep transcriptome sequencing provides new insights into the structural and functional organization of the wheat genome

LisePingault[1,2]
Email: lise.pingault@gmail.com

Frédéric Choulet[1,2]
Email: frederic.choulet@clermont.inra.fr

Adriana Alberti[3]
Email: aalberti@genoscope.cns.fr

Natasha Glover[1,2]
Email: nattyglove@gmail.com

Patrick Wincker[3,4,5]
Email: pwincker@genoscope.cns.fr

Catherine Feuillet[1,2,6]
Email: catherine.feuillet@bayer.com

Etienne Paux[1,2*]

*Corresponding author
Email: etienne.paux@clermont.inra.fr

[1]INRA UMR1095 Genetics, Diversity and Ecophysiology of Cereals, 5 chemin de Beaulieu, 63039 Clermont-Ferrand, France

[2]University Blaise Pascal UMR1095 Genetics, Diversity and Ecophysiology of Cereals, 5 chemin de Beaulieu, 63039 Clermont-Ferrand, France

[3]CEA/DSV/IG/Genoscope, 2 rue Gaston Cremieux, 91000 Evry, France

[4]CNRS UMR 8030, 2 rue Gaston Crémieux 91000 Evry, France

[5]Universitéd'Evry, CP5706 Evry, France

[6] Current address: Bayer CropScience, 3500 Paramount Parkway, Morrisville, NC 27560, USA

## ABSTRACT

### Background

Because of its size, allohexaploid nature and high repeat content, the bread wheat genome is a good model to study the impact of the genome structure on gene organization, function and regulation. However, because of the lack of a reference genome sequence, such studies have long been hampered and our knowledge of the wheat gene space is still limited. The access to the reference sequence of the wheat chromosome 3B provided us with an opportunity to study the wheat transcriptome and its relationships to genome and gene structure at a level that has never been reached before.

### Results

By combining this sequence with RNA-seq data, we constructed a fine transcriptome map of the chromosome 3B. More than 8,800 transcription sites were identified, that are distributed throughout the entire chromosome. Expression level, expression breadth, alternative splicing as well as several structural features of genes, including transcript length, number of exons and cumulative intron length were investigated. Our analysis revealed a non-monotonic relationship between gene expression and structure and led to the hypothesis that gene structure is determined by its function ("genome design" model) whereas gene expression is subject to energetic cost ("selection for economy" model). Moreover, we observed a recombination-based partitioning at the gene structure and function level.

### Conclusions

Our analysis provides new insights into the relationships between gene and genome structure and function. It revealed mechanisms conserved with other plant species as well as superimposed evolutionary forces that shaped the wheat gene space, likely participating in wheat adaptation.



### Keywords

Wheat, chromosome 3B, transcriptome, gene expression, gene structure, alternative splicing, partitioning, RNA-seq

## BACKGROUND

In angiosperms, genome size is extremely variable, ranging from 63 Mb in *Genlisea margaretae* to 148,900 Mb in *Paris japonica*, *i.e.* a 2,400-fold difference [1]. By contrast, the gene content seems relatively constant, with an average number of 30,000 and a 2- to 3-fold difference per diploid genome[2, 3]. As a consequence, the gene space organization differs strikingly from one genome to another. For example, plants with small genomes such as *Arabidopsis thaliana* (125 Mb) and *Brachypodium distachyon* (272 Mb) exhibit an even distribution of their genes along their chromosomes [4, 5] whereas for plants with intermediate size genomes such as *Populus trichocarpa* (485 Mb) and *Vitis vinifera* (487 Mb), alternation between high gene density regions and low gene density regions is observed [6, 7]. This tendency is even stronger in plants with large genomes such as *Glycine max* (1,115 Mb) and *Zea mays* (2,300 Mb) in which a positive gradient of gene density from the centromere to the telomeres has been observed [8, 9]. Beside the overall organization of genes, several studies revealed a non-random distribution of genes along chromosomes, resulting in clusters of genes sharing the same expression profile, the same function or involved in the same metabolic pathway [10-16]. In addition, relationships between gene structure and expression were reported in various organisms[17-19]. Altogether, these studies suggest a high degree of organization in gene space and interplay between genome and gene structure, function and regulation.

With 220 million hectares, bread wheat (*Triticum aestivum* L.) is the most widely grown and consumed crop worldwide providing staple food for 30% of the world population. Beside its socio-economic importance, bread wheat is also a good model for studying complex genome species. Indeed, with its large 17-Gb, allohexaploid (6x = 2*n* = 42, AABBDD) and highly repetitive (>80% transposable elements) genome, wheat is one of the most complex crop species. Other species share some of these features, but none of them, at least among cultivated species, combine the three. For example, the loblolly pine genome is the largest genome sequenced so far (22 Gb) but it is diploid [20]. Cotton is a polyploid species but has a smaller genome (2.5 Gb) [21] and so far only wild diploid relatives were sequenced [22, 23]. The maize and sorghum genomes are highly repetitive but are diploid and smaller in size [8, 9].

The wheat gene space organization and expression have been extensively investigated in the past decades. Many expression analyses have been conducted using either microarrays or RNA-Seq but most of them were aiming at deciphering specific processes, such as grain development or response to stresses(for examples, see [24-28]). Other studies aimed at

studying the gene space organization and reported on the existence of a gene gradient along the centromere –telomere axis as well as an organization of genes in small gene islandsand co-expression / co-function clusters (for examples [29-31] and references therein).However, very few of these studies really investigated the relationships between genome and gene structure and function, mainly because of the lack of a reference genome sequence. The access to physical maps of wheat chromosomes provided the first opportunities to study gene regulation with respect to their physical position [30] although there were still limited to efficiently address this question. Recently, several initiatives aimed at generating draft genome sequences of hexaploid wheat or its diploid progenitors [32-35].While they provided a quite comprehensive catalogue of wheat genes as well as novel data on gene evolution and expression, the highly fragmented nature of the sequence assemblies limits our ability to decipher the relationships between genome organization and gene regulation.

Recently, we have produced a 774-Mb reference sequence of the hexaploid wheat chromosome 3B [36]. Sequence annotation predicted 7,264 genes that were distributed along the chromosome with a gradient of density from centromere to telomeres. The distribution of structural and functional features along the chromosome revealed partitioning correlated with meiotic recombination. Three main regions were identified: two distal regions of 68 Mb (region R1; from 1 to 68 Mb) and 59 Mb (region R3; from 715 to 774 Mb) on the short and long arms, respectively and a large proximal region of 648 Mb (region R2; from 68 to 715 Mb) spanning the centromere. In addition, we delineated a 122-Mb central region (from 265 to 387 Mb), enriched in centromere-specific transposable elements, as the centromeric-pericentromeric region of chromosome 3B.

Here, we report a detailed analysis of the chromosome 3B transcriptional landscape. By combining deep transcriptome sequencing data covering the whole plant development with the reference sequence of the chromosome, we identified transcriptionally active regions distributed throughout the entire chromosome. Relationships between genome and gene structure and function revealed different mechanisms governing the gene space organization, regulation and evolution.

**Table 1.** General features of the chromosome 3B pseudomolecule transcriptionally active regions.

| **Transcriptionally active regions** | **Predicted*** | **Expressed** |
|---|---|---|
| Protein-coding genes | | |
| Total | 7,264 | 5,185 |
| Full genes | 5,326 | 4,125 |
| Pseudogenes & fragments | 1,938 | 1,060 |
| Novel transcribed regions | | |
| Total | - | 3,692 |
| Putative lincRNAs | - | 596 |
| *cis*-NATs | | 635 |

* according to [35].

## Results and discussion

### Chromosome 3B contains more than 8,800 transcriptionally active regions

To study the expression profiles of hexaploid wheat chromosome 3B genes during the life cycle of a wheat plant and establish a transcriptome atlas for this chromosome, deep transcriptome sequencing was conducted in duplicates in 15 wheat samples corresponding to five different organs (leaf, shoot, root, spike and grain) at three developmental stages each [29]. Strand-nonspecific and strand-specific libraries were used to produce2.52 billion paired-end reads (232 Gb) and 615.3 single-end reads (62Gb), respectively. The reads were then mapped to the chromosome 3B reference sequence[36], without allowing for any mismatches in order to discriminate chromosome 3B expressed genes from homoeologous and paralogous copies. Eventually, 3.66% of reads mapped onto chromosome 3B of which 98% were mapped uniquely. Ninety-five percent of the reads matched sequences annotated as genic regions whereas the remaining five percent mapped to regions where no protein-coding gene was predicted by the annotation [36].

Within the 774.4 Mb comprising the pseudomolecule of chromosome 3B, 8,877 transcriptionally active regions (TARs) were identified, corresponding to an average density of one TAR every 87 kb (Table 1). Among these, 5,185 corresponded to predicted gene models, including pseudogenes and gene fragments [36]. This represents 71.4% of the 7,264 predicted gene models. The genes contained on average 4.6 exons, ranging from one to 53, which is similar to what was found in *B. distachyon*(5.2), rice (3.8), maize (4.1), sorghum (4.3) and *Triticum urartu* (4.7) [9, 34, 37-39].The percentage of expressed genes is slightly lower than the ones reported in other plant species. Indeed, a microarray analysis of the rice transcriptome performed in seedling shoots, tillering-stage shoots and roots, heading, filling-stage panicles and suspension-cultured cells detected expression for 86% of the 41,754 known and predicted gene models present on the microarray[40]. More recently, Lu *et al.*[41]conducted an RNA-Seq analysis on seeds from three rice cultivated subspecies and found that 83.1% of the 46,472 annotated gene models were expressed. Similarly, in maize, microarray-based transcript profiling in 60 distinct tissues representing 11 major organ systems revealed that 91.4% of the genes were expressed in at least one tissue [42]. More recently, Sekhon and coworkers [43] performed RNA-Seq experiments on a subset of 18 selected tissues representing five organs and showed that 74.7% of the 39,429 genes from the filtered gene set were transcribed. In soybean RNA-Seq analysis revealed that 80.4% of 69,145 putative genes are expressed in a least one of the 14 tissues analyzed[44].The lower percentage of genes expressed in wheat might be the evidence for

the early step of gene loss after polyploidization or suggest a small impact of polyploïdisation on gene silencing. This is consistent with previous studies conducted in newly synthesized polyploid wheat and rapeseed where 7.7 and 4.1% of the sequences showed alteration in gene expression [45, 46]. To estimate the exact extent of gene silencing in hexaploid wheat, a comparison with diploid and tetraploid progenitors would be required. However, when considering only genes likely to be functional (hereafter referred to as "full genes"), the percentage of expressed genes rose to 77.5% (4,125 / 5,236), which is similar to the percentages found in maize and soybean using a similar number of conditions [43, 44]. Beside full genes, 54.7% (1,060 / 1,938) regions annotated as pseudogenes or gene fragments in the pseudomolecule were found to be expressed in at least one condition. In other species such as *A. thaliana* and rice, EST analyses revealed expression for 2-5% and 2-3% pseudogenes, respectively[47].Another study conducted on 1,439 rice pseudogenes using Massively Parallel Signature Sequencing tags suggested that up to 12% are expressed in at least one of the 22 samples studied [48]. These proportions strongly differ from our results. One cannot exclude that the percentage of pseudogenes expressed on chromosome 3B could be overestimated as a result of the RNA-Seq technology that cannot completely discriminate pseudogene expression from close functional copies that might be present elsewhere in the genome. In an attempt to assess this overestimation, we searched the recently released draft assembly of the wheat genome[32] for additional copies of pseudogenes in the genome. Overall, 511 out of 1,060 (48.2%) had at least one other copy, whereas 51.8% were found to be present in one single copy located on chromosome 3B. Assuming that for the 48.2% “multicopy” pseudogenes, transcripts were not produced by the 3B loci, our results suggest that 28% of the chromosome 3B pseudogenes are still expressed, which is much higher than what has been observed in other organisms so far. Transposable elements (TEs) have been shown to be able to generate sense or antisense transcripts of adjacent genes [49]. Given the high proportion (> 85%) of the wheat genome covered by TEs, one can hypothesize that some TEs provide a promoter for transcription of adjacent pseudogenes. In addition, while they have long been considered as non-functional units, several studies suggest that pseudogenes might play a role in regulation through antisense regulation of their parental gene, competition for miRNA, generation of small-interfering RNA or production of short proteins or peptides [50, 51]. The high percentage of expressed pseudogenes found in wheat compared to other species might therefore be due to their role in the regulation of homoeologous or paralogous gene expression.

For 28.6% of the predicted gene models (2,079), we failed to detect any expression. This result probably reflects the fact that these genes might be expressed in specific conditions that have not been studied in the present work. Indeed, a Gene Ontology term analysis of

these non-expressed genes revealed enrichment in biological processes such as "gametophyte development" or "response to temperature stimulus". In addition, 57.5% of the non-expressed genes are non-syntenic with *B. distachyon*, rice and sorghum, suggesting that some of these genes might have been duplicated and translocated without their regulatory sequences, leading to their transcriptional inactivity. Finally, it is worth noting that the proportion of non-expressed pseudogenes is twice as high as the proportion of non-expressed functional genes (45.3% *vs.* 22.6%). As a result, the distribution pattern of non-expressed genes along the chromosome was found to be highly correlated with that of pseudogenes ($r_S$= 0.81, $p$< 2.2e10$^{-16}$).

In addition to the predicted gene models, expression was detected for 3,692 loci in unannotated regions. These so-called novel transcribed regions (NTRs) represented on average 22% of all TARs. Twenty-eight percent (1,033 / 3,692)of these NTR-translated sequences shared weak similarity with plant proteins, mainly TE-encoded proteins or hypothetical proteins and might therefore be protein-coding genes (or pseudogenes). Out of the 2,659with no similarity with plant proteins, 596 were longer than 200 nt and did not carry ORFs longer than 300 AA. These NTRs might therefore correspond to long intergenic non-coding RNAs (lincRNAs) as defined by Liu and coworkers [52].Based on this number, one could speculate that roughly 10,000 lincRNAs should be expressed in the whole wheat genome, or 3,300 per diploid genome. This number is comparable to thatof expressed lincRNAs reported *A. thaliana*[52]and poplar [53](2,708 and 2,542, respectively). Out of these 596 putative lincRNAs, 91.1% and 93% were found in the *Triticum urartu* and *Aegilops tauschii* genomes, respectively. The percentage decreases to 69.3% when looking at the barley genome. An even more drastic drop was observed when moving out of the Triticeae tribe, with only 14.8%, 7% and 6.2% of the putative lincRNAs conserved in the *B. distachyon*, rice and sorghum genomes, respectively. These findings suggest that most of these putative lincRNAs are functional elements that have been acquired by the wheat and more largely the Triticeae genomes in the time course of their evolution.

Beside lincRNAs that are located in intergenic regions and therefore do not overlap with protein-coding genes, *cis*-natural antisense transcripts (NATs) are another form of long non-coding RNAs [54-56]. To estimate the extent of *cis*-NATs in wheat, oriented RNA-Seq libraries from the five organs were constructed and reads were mapped on chromosome 3B without allowing mismatches. Out of the 5,185 expressed genes, 635 (12.2%) were found to be transcribed on the reverse strand as well, therefore producing a *cis*-NAT. It is worth noting that *cis*-NATs originate preferentially from syntenic genes (72.4%) and the vast majority (84.9%) concerned full genes. A previous study conducted in wheat using microarray identified 110 NATs at the whole genome level[57]. Conversely, Serial Analysis of Gene

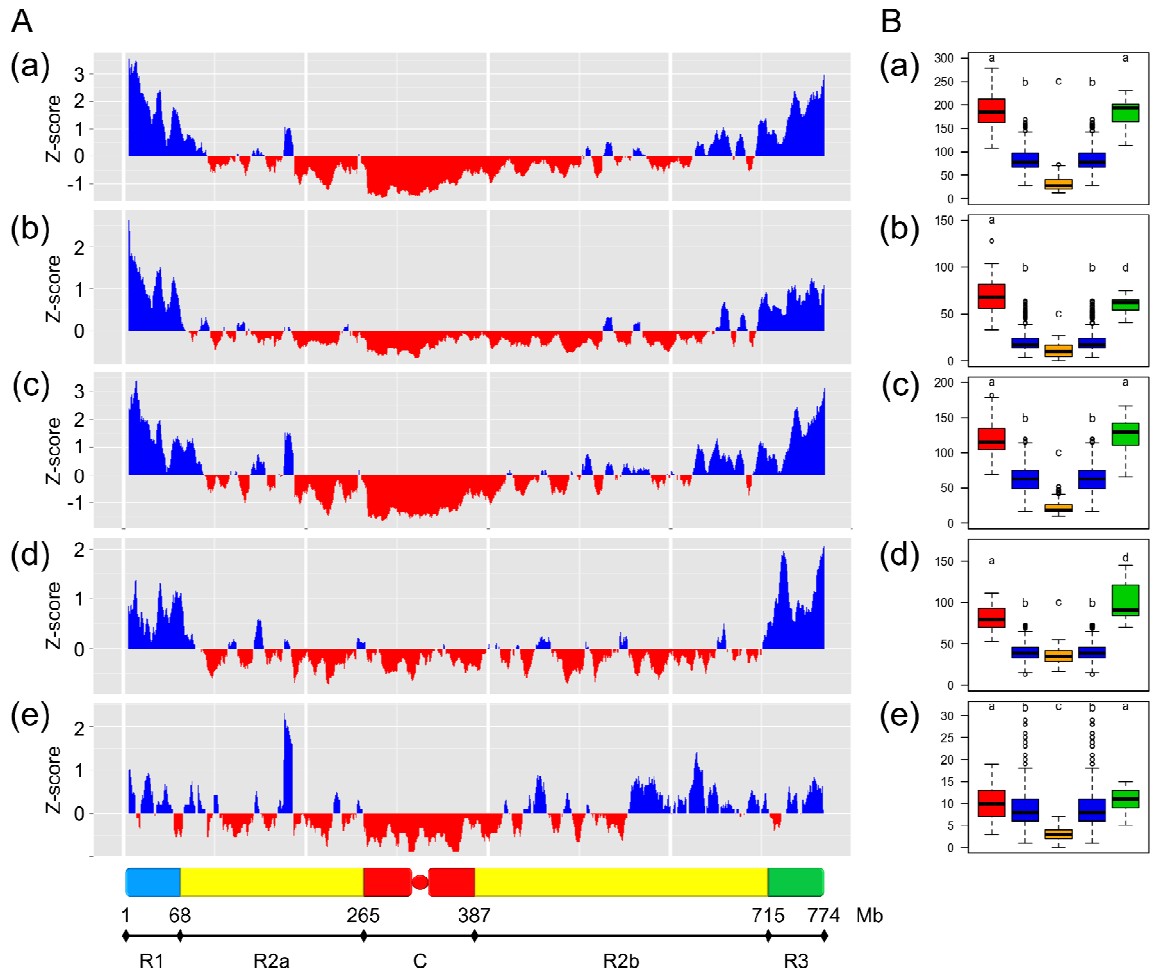



**Figure 1.**Distributionof the functional regions on the chromosome 3B.
**A**: Z-score of the density of features in a 10-Mb sliding window (step 1 Mb) along chromosome 3B. Positive values are in blue, negative values in red. (a) predicted genes; (b) nonexpressed predicted genes; (c) expressed genes; (d) NTRs; (e) *cis*-NATs. The five main regions of the chromosome 3B are depicted at the bottom of the graph: R1 and R3 in blue; R2a and R2b in yellow; C in red. The borders of these regions are indicated in Mb.
**B**: boxplots of the density of features in the five main regions of the chromosome 3B (R1 and R3 in blue; R2a and R2b in yellow; C in red). (a) predicted genes; (b) nonexpressed predicted genes; (c) expressed genes; (d) NTRs; (e) *cis*-NATs.

Expression showed that up to 25.7% of wheat were represented by reverse tags [58]. Such widespread occurrence of antisense transcription has already been reported in other plant species such as *A. thaliana*, rice or maize where 2.8 to 9.7% of genes produce antisense transcripts [56, 59, 60]. *Cis*-NATs can regulate gene expression at the transcriptional or post-transcriptional level through various mechanisms [54, 61]. In a polyploid species, one can hypothesize that they play a role in the regulation of homoeologous copies.

**Transcription sites are distributed throughout the entire chromosome 3B**

The distribution of predicted protein-coding genes, non-expressed genes, expressed genes, NTRs and cis-NATs density was analyzed along chromosome 3B (Figure 1A). In addition, we computed these densities in five regions modified from Choulet *et al.*[36] based on both recombination and (centromere-specific) TE content segmentation analyses: R1 (1-68 Mb), R2a (68-265 Mb), C (265-387 Mb), R2b (387-715 Mb) and R3 (715-774 Mb) (Figure 1B). The density of expressed genes was highly correlated with the distance to the centromere ($r_s$ = 0.77,*p*<2.2e-16) and was found to follow that of predicted protein-coding genes ($\chi^2$ test = 415.84, df= 762). The overall average density was 6.5 ± 3.3 genes / Mb, ranging from 1.0 in the centromeric region up to 18.2 at the most telomeric end of the short arm. With an average density of 4.8 ± 4.1per Mb, NTRs were slightly less abundant than expressed protein-coding genes but followed the overall gene distribution. However, their proportion was found to be much higher in the pericentromeric C region. Since this region corresponds to the part of the chromosome where the TE density is the highest, this suggests that some of these NTRs might actually be transcribed from adjacent TE promoters. Whether these RNAs are 'transcriptional' noise or have a biological function remains to be investigated. The distribution of *cis*-NATs is slightly more even along the chromosome, suggesting that proximal genes are more prone to antisense transcription than distal ones. Once again, this might be due to the high abundance of TEs in these regions that would provide promoters for the transcription of adjacent genes.

Taken together, these results clearly demonstrate that transcription occurs all along chromosome 3B and is not restricted to distal regions. This is in complete agreement with our previous analyses using microarray hybridizations of BAC pools and mRNA samples [30] and with observations from Abranches *et al* [62]who demonstrated that active transcription sites are distributed throughout the wheat genome and do not show any preferential localization in the nuclei. More recently, Baker *et al.*[63] provided evidences that genes located in the low recombining pericentromeric regions were expressed at a level that was similar to that of

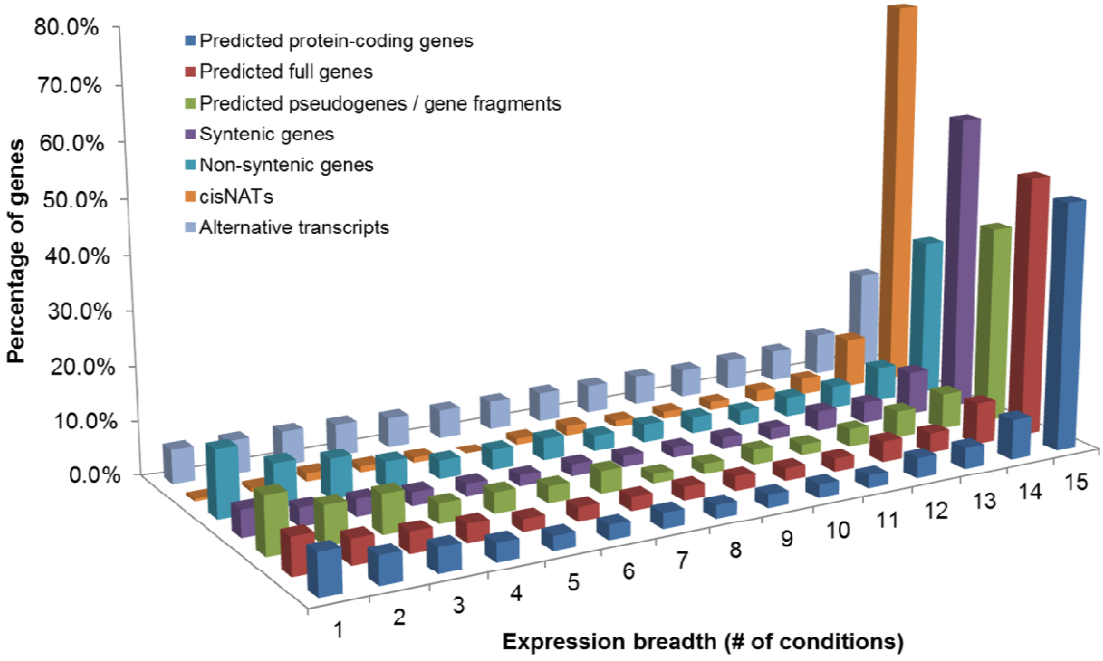



**Figure 2.**Distribution of the percentage of transcriptionally active regions expressed in the different number of experimental conditions.
Regions were classified according to their expression breadth, *i.e.* the number of conditions in which they were expressed, from 1 to 15. Dark blue: predicted protein-coding genes; red: predicted full genes; green: predicted pseudogenes / gene fragments; purple: syntenic genes; cyan: non-syntenic genes; orange: *cis*-NATs; light blue: alternative transcripts.

genes in high recombining distal regions in barley. Thus, while gene density follows an increasing gradient along the centromere-telomere axis that correlates with recombination, this distribution does not seem to relate to the overall transcription capacities of wheat genes.

**Expression level, expression breadth and alternative splicing are correlated**

The number of expressed genes was found to be comparable across the 15 conditions, with on average 3,734 ± 228 genes expressed per condition. A similar trend was observed in other species such as maize [42], soybean[44] or peach [64]. The average expression breadth (*i.e.* the number of conditions in which a gene is expressed)for the 5,185 expressed gene models was 10.8, with 46.2% (2,396) of the genes expressed in all conditions and 7.6% (396) exhibiting a condition-specific expression profile. At the organ level, the number of organ-specific genes ranged from 77 in leaf to 243 in spike. These proportions of condition-specific genes are not similar for all types of genes (Figure 2). For example, pseudogenes and gene fragments were found to be more specific than full genes with only 36.9% of them being expressed in 15 conditions and 10.7% in one single condition (*vs.* 48.6% and 6.9% for full genes, respectively). A similar trend was observed when comparing syntenic and nonsyntenic genes that were identified by comparative analysis along the 3B sequence [36]. Indeed,29.6% of the nonsyntenic genes were expressed in 15 conditions and 12.5% in one single condition, whereas 55.5% and 4.9% of syntenic genes were found to be expressed in 15 and one condition, respectively. By contrast, 73.7% of genes showing anti-sense transcription were expressed in 15 conditions while only very few of them (0.5%) were specific to one single condition. This reinforces the idea that *cis*-NATs serve as post-transcriptional regulators of gene expression [65-67].

Expression breadth was found to be correlated with expression level. This correlation is not unexpected since genes that are widely expressed such as house-keeping genes tend to show a higher expression level than condition-specific genes [18, 68, 69]. However, to some extent, one cannot exclude that this relationship between expression level and expression breadth reflects the fact that expression is not detected in some conditions and that some condition-specific genes might just be low expression genes.

Our analysis revealed 30,232 transcripts originating from the 5,185 chromosome 3B expressed genes. Thirty nine percent of the genes were transcribed in one single mRNA in our conditions whereas splicing variants were detected for 61.4%, with an average of 5.8 alternative transcripts per gene. When considering multiexonic genes only, the percentage of alternatively spliced genes raised to 75.4%. While alternative splicing(AS) is a general phenomenon in plants, the overall AS level differs strikingly between species. Indeed, previous studies reported that 61% and 48% of *A. thaliana* and rice genes undergo AS,

respectively[41, 70], whereas only 6.3% and 15.9% of expressed genes are under the potential influence of AS in *B. distachyon*and soybean, respectively[71, 72].In barley, 55% of high confidence genes and 73% of intron-containing high confidence genes have evidence of AS [73]. This high similarity between barley and wheat, as well as differences with that of rice and *B. distachyon* suggests that the AS level might have evolved differently in grasses. Conversely, considering that the level of AS observed in wheat was similar to that of *A. thaliana*, it is very unlikely that these differences between species are linked to genome size and complexity. However, one cannot exclude that differences originate from experimental design. In*A. thaliana*, the predicted AS level increased from 1.2 to 61% between 2003 and 2012, mainly as a result of the advent of high-throughput technologies [74].In addition, alternative transcripts have been hypothesized to be tissue- or condition specific [75]. As our results are based on the study of plants grown in normal conditions we cannot exclude that the percentage of AS genes is underestimated and might increase with the inclusion of other samples such as plants grown under stress conditions.

Beside the differences observed in the overall AS level between species, we found differences in the relative abundance of the main types of AS, namely exon skipping (ES), alternative splice sites (A3SS and A5SS), intron retention (IR) and mutually exclusive exons (MXE) [76]. In wheat, IR was found to be the predominant type, with 35% of all events, followed by A3SS (27%), ES (21%), A5SS (16%) and MXE (0.9%). In *A. thaliana*, rice and *B. distachyon*, IR accounts for more than 50%[71, 77] whereas the predominant type was found to be ES in the peach genome, with 43% of all observed events[64]. Such differences strongly reinforce the idea that, despite the fact that AS is a common phenomenon shared by most if not all plant species, specificities have been acquired by the different plant species during the course of their evolution.

While 46.2% of the 5,185 genes were expressed in 15 conditions, only 18.6% of 30,232 transcripts appeared to be present in all conditions, which is very similar to what has been observed in barley [73]. In addition, 95% of the AS transcripts originating from the same gene exhibited different expression profiles, as revealed by a hierarchical clustering of the 30,232 transcripts (data not shown). As a consequence, the number of alternative transcripts was found to be positively correlated with the expression breadth. These findings strongly suggest that AS variants have complementary functions across organs or developmental stages.

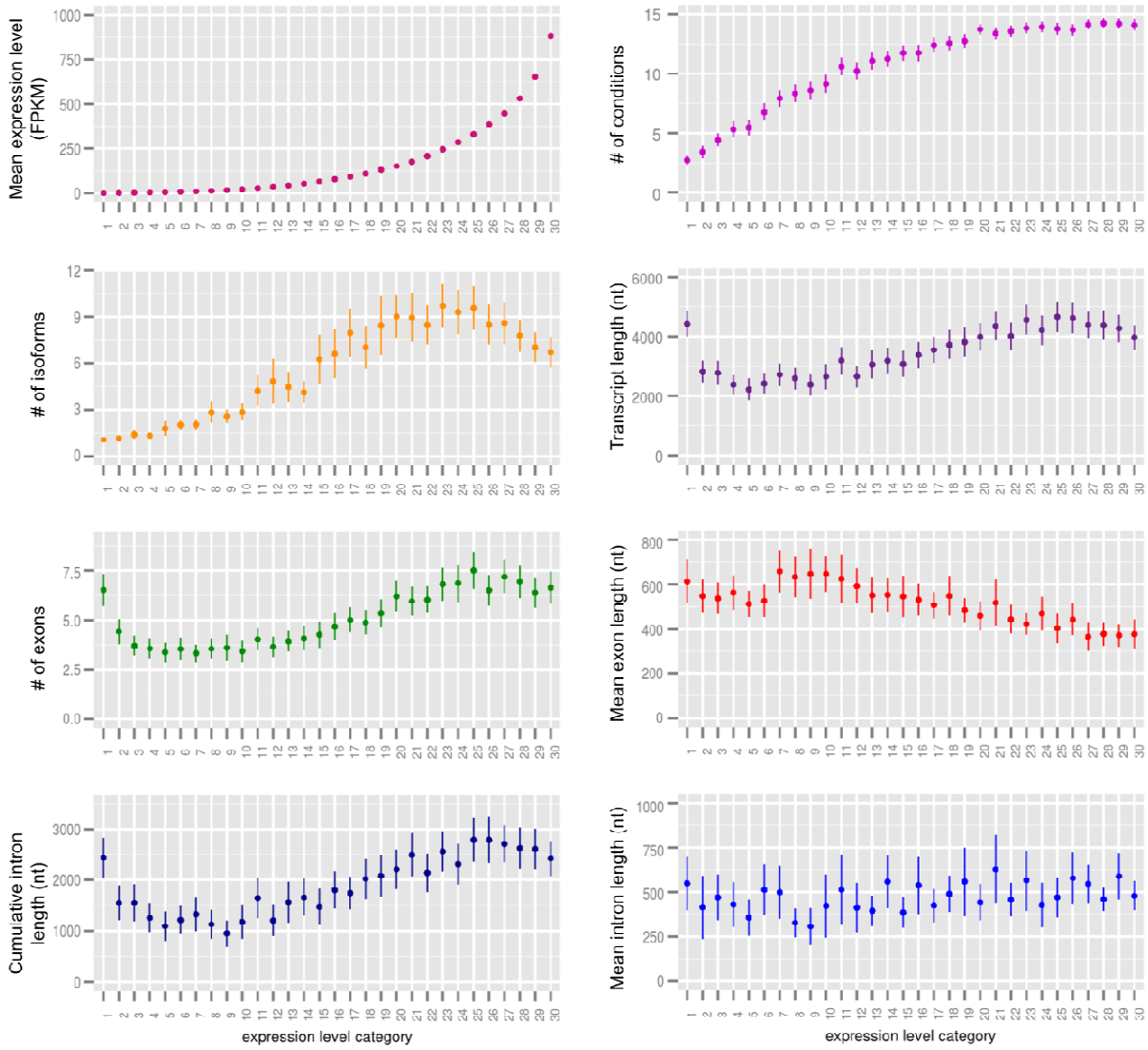



**Figure 3.**Relationships between gene expression and gene structural and functional features.
Expression levels are binned into 30 categories. Each dot is the mean value for genes in the given expression category, and the error bar indicates the standard deviation of the mean.

**A non-monotonic relationship between gene expression and gene structure**

A negative correlation was observed between the transcript size and the expression breadth, with shorter transcripts being expressed in more conditions. This is consistent with previous studies indicating that house-keeping genes which are expressed in more conditions are generally more compact than genes expressed in specific conditions[18, 68, 69, 78]. Such findings could be explained by the "selection for economy" model [18, 79]. In this model transcription and translation are both time- and energy consuming and, as a consequence, widely and highly expressed genes tend to be more compact to reduce the energetic cost [80, 81].

We then investigated the correlation between expression level and gene structure, in terms of transcript size, number of exons, cumulative intron length, mean exon and intron length, and number of alternative transcripts. To this aim, genes were grouped in 30 classes of similar size based on their average expression level (*i.e.* the average FPKM value in a number of conditions where the gene is expressed), as done by Carmel and Koonin [17]. Then, the average values of different variables in each of the 30 expression level classes were computed across the genes (Figure 3).

For the transcript size, the number of exons, the cumulative intron length and the number of alternative transcripts, non-monotonic relationships were found with the expression level, resulting in an approximate bell-shaped dependence (Figure 3). For all features, the area of the inflexion point was comprised between classes 20 and 25 which is also the area where expression breadth reaches a plateau. Following criteria defined by Hansey *et al.*[82], genes in categories 1 to 4 correspond to low expression genes (mean expression level < 5 FPKM), those in categories 5 to 21 to medium expression genes (5 ≤ mean expression level < 200 FPKM) and those in categories 22 to 30, to high expression genes (mean expression level ≥ 200 FPKM). Interestingly the inflexion point also corresponded to the threshold between medium and high expression genes. For medium expression genes, the expression level was positively correlated with the structural features whereas for high expression genes, we found a negative correlation. For low expression genes, the observed relationship might be an artifact resulting from the detection threshold of low abundance transcripts. Indeed some of these genes might have been considered as expressed when actually they were not. Therefore, these four classes (1-4) might not be reliable as they might contain a mix of expressed and non-expressed genes leading to average structural feature values that are not representative of expressed genes. For the mean exon and intron length, no clear relationship was observed even though the mean exon length tends to decrease as the expression level increases. Such a non-monotonic relationship has already been observed in

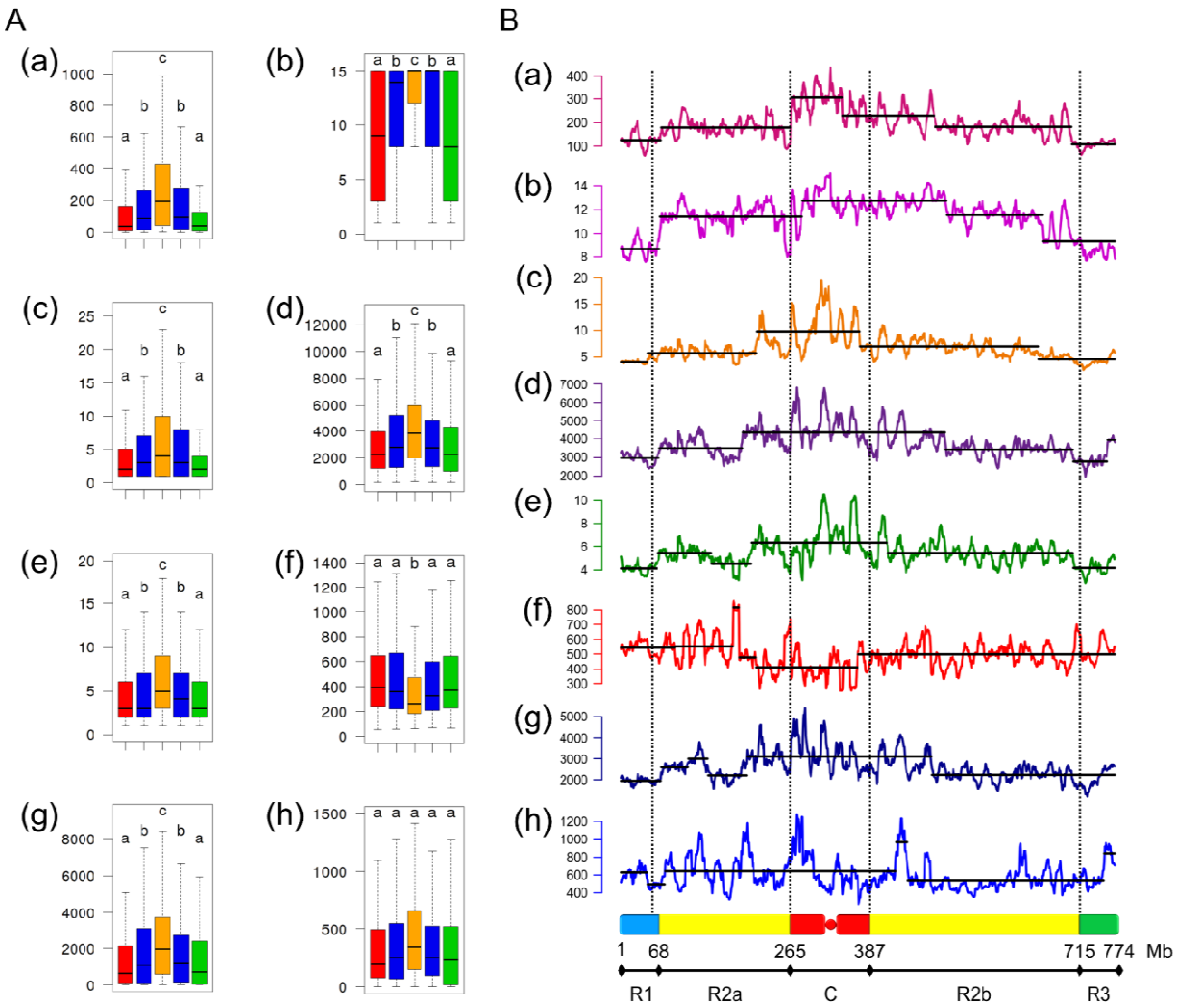



**Figure 4.**Distribution and functional partitioning of wheat chromosome 3B.
**A**:boxplots of structural and functional gene features in the five main regions of the chromosome 3B (R1 and R3 in blue; R2a and R2b in yellow; C in red). (a) gene expression in FPKM; (b) expression breadth in number of conditions; (c) number of alternative transcripts per gene; (d) transcript length; (e) exon number; (f) mean exon length; (g) cumulative intron length; (h) mean intron length.
**B**: Distribution and segmentation analysis of (a)gene expression in FPKM, (b) expression breadth in number of conditions, (c) number of alternative transcripts per gene, (d) transcript length, (e) exon number, (f) mean exon length, (g) cumulative intron length, (h) mean intron length. Sliding window size: 10 Mb, step: 1 Mb.

other organisms, including human, *Caenorhabditis elegans*, *Drosophila melanogaster*, *A. thaliana* and soybean [17, 19]. If the "selection for economy" fits for highly expressed genes, it cannot apply to low to medium expression genes. The "genome design" model has been proposed to explain this relationship [18, 69, 78, 79]. It suggests that the structural features of a gene are mostly determined by its functional load. Highly and widely expressed genes would not require a fine regulation and therefore less regulatory sequences. By contrast, for low / medium expression, condition-specific genes, longer intragenic non-coding sequences would allow for a more complex regulation. Since the number of alternative transcripts follows the same distribution, one can hypothesize that the greater number of exons and the larger intronic sequences might allow for a greater transcriptional complexity leading to a greater specificity in gene expression. A detailed analysis of transcript size based on the 30,232 isoforms showed a negative correlation with expression level, regardless of the expression class (Figure S1 in Additional file 2). This finding reinforces the hypothesis that gene structure would be determined by its function ("genome design" model) whereas the expression of the different transcripts from a given gene would be subject to the energetic cost ("selection for economy" model).

**Gene structural and functional features are partitioned on chromosome 3B**

To investigate their relationship with chromosome partitioning, the expression breadth, expression level, transcript size, number of exons, cumulative intron length, mean exon and intron length, and the number of alternative transcripts were computed in the five regions of chromosome 3B, namely R1, R2a, C, R2b and R3 (Figures 4A). For all features but mean exon and intron length, the regions can be classified in three contrasting groups. The first one includes regions R1 and R3, the second one, R2a and R2b and the third one, C. All of the features decrease along the centromere – telomeres. Thus, on average, genes in distal regions are expressed at a lower level, more specifically and have fewer isoforms than those in the proximal regions. In addition, they are shorter, with fewer exons and shorter intronic sequences. Genes located in region C tend to have shorter exons, while for the mean intron length, no significant differences were observed between the five regions. A segmentation analysis of these properties suggests a partitioning of the chromosome rather than a regular gradient from centromere to telomeres (Figure 4B). Interestingly, the boundaries of the distal segments fit almost perfectly with the R1 and R3 regions defined by Choulet *et al.*[36] based on recombination.

To see to what extent the non-monotonic relationship between expression level and gene structural features observed at the whole chromosome level is conserved at the region level,

we applied the same analysis to regions R1/R3 and R2a/R2b. Region C was not included due to the limited number of genes present in this region. Interestingly, the chromosomal pattern remained the same in each region (Figures S2 and S3in Additional file 1). Even though the average expression level was lower in R1/R3 regions, the mean expression level value of the inflexion point was conserved in the two regions, around 200FPKM, the approximate threshold between medium and high expression genes. This finding clearly shows that the "selection for economy" and "genome design" models apply all along the chromosome independently of other features and strongly suggests that the evolutionary forces that have led to the chromosome partitioning are distinct from the molecular mechanisms governing gene expression.

**Chromosome conformation may play a role in gene regulation**

A hierarchical clustering of the 5,185 expressed protein-coding gene primary transcripts was performed based on their expression profiles in the 15 conditions. Genes were aggregated into 55 distinct clusters according to their expression profiles. Based on the median value of intergenic distances of 30 kb, Choulet *et al.*[36] estimated that 73% of genes were organized in small islands or "*insulae*". Using the same criteria, 3,465 out of the 5,185 expressed genes (67%) were found to be organized in 1,199 *insulae*, comprising 2.9 genes on average. Out of these 3,465 genes, 1,218 (35.2%) belong to the same expression cluster as their direct neighbor, defining 718 co-expressed gene pairs. This proportion is higher than the previously reported value of 11% [30]most probably because of the higher resolution achieved with a reference sequence compared to a partial gene dataset. Such enrichment has already been reported in other organisms such as human, mouse, *A. thaliana,* rice and fruit fly where the percentage of adjacent co-expressed genes ranges from 2% to 20%[12, 13, 16, 83, 84]. However, these percentages are relatively low compared to that of wheat. One can hypothesize that the higher proportion of co-expressed genes found in wheat might result from the high rate of tandem duplication in this genome [36]. However, of the 718 co-expressed gene pairs, only 46 (6.4%) correspond to duplicated genes. This clearly shows that duplicate genes alone do not explain the observed levels of co-expression, as already reported in other organisms [10, 14, 16, 85]. Several other mechanisms have been proposed to explain the co-expression of neighboring genes, including shared promoters and chromatin remodeling. In *A. thaliana*, Chen *et al.*[85] showed that co-expression was strongly enhanced for divergently transcribed genes within a 400-bp gene distance, probably as a result of shared promoters. For longer intergenic distances, co-expression is likely mediated by shared chromatin environments. On chromosome 3B, the average intergenic distance

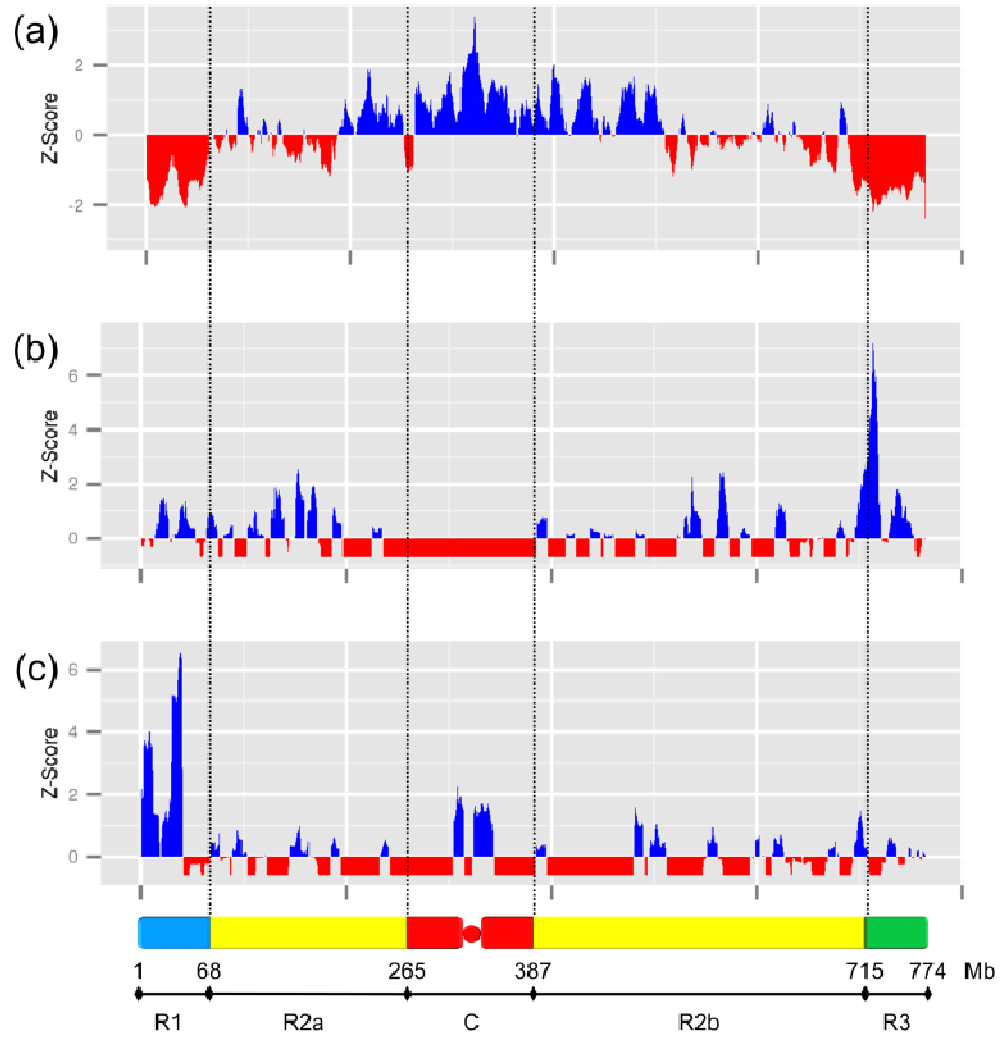



**Figure 5.**Distribution of the percentage of genes from three different expression clusters. Z-score of the percentage of expressed genes for a given cluster in a 10-Mb sliding window (step 1 Mb) along chromosome 3B. Positive values are in blue, negative values in red. (a) Constitutively expressed genes; (b) spike- and grain-specific genes; (c) genes preferentially expressed in leaves. The five main regions of the chromosome 3B are depicted at the bottom of the graph: R1 and R3 in blue; R2a and R2b in yellow; C in red. The borders of these regions are indicated in Mb.

between co-expressed neighbor genes is 6.3 kb and only 133 out of the 718 gene pairs are transcribed divergently. This suggests that shared promoters are not the main mechanism controlling the co-expression of neighbor genes and that other mechanisms such as chromatin conformation might be involved. This hypothesis is reinforced by the significant differences observed for 23 out of the 55 expression clusters between the five regions (Table S1 in Additional file 1). For example, the vast majority (63%) of the genes present in the region C belong to cluster I that correspond to genes expressed in all conditions whereas this cluster represents only 23-24% of region R1 and R3 genes (Figures 5). Region R1 is enriched in genes preferentially expressed in leaf compared to other regions. Region R3 displays a higher proportion of spike- and grain-specific genes. In addition, expression clusters oscillated along chromosome, forming chromosomal domains. These findings are consistent with the Gene Ontology term enrichment analysis that revealed that distal regions where enriched in genes involved in adaptive processes such as response to abiotic stimuli or stress[36].

Even though transcription sites are distributed throughout the entire chromosome when looking at the plant development at a whole, our results show that 3B is organized in chromosomal domains, suggesting that gene positions influences the spatio-temporal regulation of their expression. Such domains have recently been reported for genes expressed in wheat endosperm [24]. It has been shown that the spatial organization of genomes in the interphase cell nucleus is tissue-specific [86]. This positioning of chromosomes is non-random and is likely to play a role in gene regulation [87, 88]. In wheat, the interphase chromosomes are not fully decondensed but adopt a regular Rabl configuration, a highly polarized pattern with the two chromosome arms lying next to each other and the centromeres and telomeres located at opposite poles of the nuclei [89-91]. The presence of this organization is also known to vary greatly among tissues or developmental stages of an organism [90]. Then, one can hypothesize that this configuration might play a role in gene regulation through the partial decondensation of given chromosomal regions in specific tissues and at specific developmental stages, leading to the observed spatial partitioning of genes displaying similar expression profiles. This hypothesis is well supported by our previous results [36]. Indeed, a similar recombination- and expression breadth-based partitioning was found in barley in which the Rabl configuration is also observed, but not in maize which displays neither entirely Rabl nor entirely random chromosome organization [89, 90].

## CONCLUSIONS

By combining the first reference sequence of a wheat chromosome with deep transcriptome sequencing data covering the whole plant development, we constructed a high density transcription map of the wheat chromosome 3B, comprising more than 8,800 transcriptionally active regions distributed throughout the entire chromosome. By studying the relationships between genome and gene structure and expression, we unraveled two interconnected mechanisms. The first one is a universal mechanism that relates to the “selection for economy” and “genome design” models and links gene structure and function, regardless of the gene position. The second one is an evolutionary force that links gene structure and function to gene position, leading to a strong partitioning of the wheat chromosome 3B. Since this partitioning is also observed in barley but not in other grasses, one can hypothesize that it has evolved with genome organization and is related to Triticeae-specific adaptation.

## MATERIAL AND METHODS

### Sample preparation and sequencing

Total RNAs were extracted in duplicates from five organs (root, leaf, stem, spike, and grain) at three developmental stages each from hexaploid wheat *cv.* Chinese Spring [29] (Table S2in Additional file 1). RNA quality was assessed using an RNA nano Chip on the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara) and the RNA integrity number (RIN) was calculated for each sample. Only sample with a RIN greater than 7 were used for the library construction.

The 30 strand-non specific RNA-seq libraries (representing the 15 conditions in duplicates) were constructed from 4µg of total RNA using the Illumina TruSeq$^{TM}$ RNA sample preparation Kit (Illumina, San Diego, CA) according to the manufacturer's protocol, with a library insert size of 300bp (fragmentation time of 12min). Library profiles were evaluated using an Agilent 2100 Bioanalyzer. Illumina indexes were used to pool two samples per lane. Libraries were sequenced on an Illumina HiSeq2000 with 2 x 100-bp paired-end reads.

For strand-specific RNA-seq libraries, 12 µg of total RNAs from the same organ were pooled (4 µg per developmental stage) and polyA+ enriched RNAs were purified using the Ambion MicropolyA Purist Kit (Life Technologies, Carlsbad, CA). Fifty nanograms of purified poly A+ RNAs were used to construct the oriented RNA-seq libraries with the ScriptSeq v2 RNA-Seq Library Preparation kit (Epicentre, Madison, WI) following the manufacturer's instructions. After cDNA synthesis, 15 cycles PCR were performed to amplify the fragments. Libraries were purified by Ampure beads (Beckmann Coulter, Indianapolis, IN) and then quantified using a Qubit Fluorometer (Life technologies). Library profiles were evaluated using an Agilent 2100 Bioanalyzer. Each library was sequenced using 101 base-length read chemistry on one lane of a single-end (SE) flow cell on the Illumina HiSeq2000.Read quality was checked with the FastQCv0.10.0 software[92]. RNA-Seq data have been deposited under accession number ERP004714.

### Read alignment and expression analysis

Illumina reads were mapped on the chromosome 3B scaffolds using Tophat2 v2.0.8 [93, 94] and bowtie2 [95] with the default parameters except: 0 mismatch, 0 splice-mismatch. PCR duplicates were removed with Samtools [96] rmdup option and an annotation-guided read alignment was performed with Cufflinks v2.1.1 [93, 97] to reconstruct transcripts and estimate transcript abundance in units of fragments per kb of exon per million mapped reads (FPKM) [98]. Regions with FPKM values higher than zero were considered as expressed. TriAnnot-predicted regions were distinguished from unannotated regions (novel transcribed

regions, NTRs) using the –g option. NTRs were reconstructed and ORFs were detected using transcripts_to_best_scoring_ORFs.pl (Trinity) [99] and blasted against the Magnoliophyta database (BLASTX, e-value 10e-5). Based on the FPKM scale defined in by Hansey and collaborators [82] expressed genes can be divided in three classes: genes with a FPKM value below 5 are low expressed, genes with a FPKM value greater or equal to 5 and less than or equal to 200 are medium expressed, and genes with a FPKM value greater than 200 are high expressed (semi-quantitative organization).

Sequences and annotations of the reference pseudomolecule and unassigned scaffolds have been deposited in ENA (project PRJEB4376) under accession numbers HG670306 and CBUC010000001 to CBUC010001450, respectively.

**Segmentation / change-point analysis**

Segmentation analyses were performed using the R package changepoint v1.0.6 [100] with Binary Segmentation method and BIC penalty on the mean change. The different features that were subjected to this analysis were: recombination rate, transposable element density, predicted gene density, number of condition in which a gene is expressed. All these features were calculated in sliding windows of 10 Mb with a step of 1 Mb.

**Statistical analysis**

All statistical analyses were performed with the R software[101]. Shapiro-Wilk test was used to test for normality of distribution. Correlation analyses were performed with Spearman rank correlation method. Outlier detection was performed using the formula: (Quantile3-Quantile1) x 3 / Quantile3, based on FPKM value and transcripts length of each gene. Genes were classified according to their average expression level and divided in 30 classes, with the same number of genes per class. R package ggplot2 was used to draw plot. Average comparison was performed using Welsh t.test to test for statistical significance between the 5 regions.

**Hierarchical clustering**

Hierarchical clustering was performed using the Hierarchical Clustering Explorer 3.5 software (http://www.cs.umd.edu/hcil/hce/) with the complete linkage method and the Pearson correlation coefficient. The minimal similarity to establish the clusters was set to 0.641 which is the Pearson correlation significant at the P-value threshold of 0.01.

## LIST OF ABBREVIATIONS USED

A3SS: alternative 3' splice site

A5SS:alternative 5' splice site

AS: alternative splicing

BLAST: Basic Local Alignment Search Tool

bp: base pairs

*cis*-NAT:*cis*-natural antisense transcript

DNA: deoxyribonucleic acid

FPKM: fragments per kb of exon per million mapped reads

IR: intron retention

kb: kilo base pairs

lincRNA: long intergenic non coding ribonucleic acid

Mb: megabase pairs

miRNA: micro ribonucleic acid

mRNA: messenger ribonucleic acid

MXE: mutually exclusive exon

NAT: natural antisense transcript

ncRNA: noncoding ribonucleic acid

nt: nucleotide

NTR: novel transcribed region

ORF: open reading frame

RNA : ribonucleic acid

RNA-Seq: sequencing of RNA

TE: transposable element

TAR : transcriptionally active region

## COMPETING INTERESTS

The authors declare that they have no competing interests.

## AUTHORS' CONTRIBUTIONS

LP constructed RNA-Seq libraries, carried out RNA-Seq data analyses and drafted the manuscript. FC participated in RNA-Seq data analyses and interpretation. AA and PW produced RNA-Seq data. NG performed paralogous and syntenic / nonsyntenic gene analyses. CF acquired the funding and corrected the manuscript. EP designed research, supervised all the analyses and drafted the manuscript. All authors read and approved the final manuscript.

## ACKNOWLEDGEMENTS

The authors would like to thank Hélène Rimbert for assistance in bioinformatics analysis. They also thank the International Wheat Genome Sequencing Consortium (IWGSC) for pre-publication access to the chromosome-based draft sequence of the wheat genome. This work was supported by grants from the French National Research Agency (ANR-09-GENM-025 3BSEQ) and France Agrimer. LP was funded by a grant from Région Auvergne.

## REFERENCES


1. Bennett MD, Leitch IJ: **Nuclear DNA amounts in angiosperms: targets, trends and tomorrow**. *Ann Bot* 2011, **107**:467-590.
2. Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL, Salzberg SL, Feng L, Jones MR, Skelton RL, Murray JE, Chen C, Qian W, Shen J, Du P, Eustice M, Tong E, Tang H, Lyons E, Paull RE, Michael TP, Wall K, Rice DW, Albert H, Wang ML, Zhu YJ *et al.*: **The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus)**. *Nature* 2008, **452**:991-996.
3. **CoGePedia** [http://genomevolution.org/wiki/index.php/Main_Page]
4. The Arabidopsis Genome Initiative: **Analysis of the genome sequence of the flowering plant Arabidopsis thaliana**. *Nature* 2000, **408**:796-814.
5. The International Brachypodium Initiative: **Genome sequencing and analysis of the model grass Brachypodium distachyon**. *Nature* 2010, **463**:763-768.
6. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyere C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V *et al.*: **The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla**. *Nature* 2007, **449**:463-467.
7. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q *et al.*: **The genome of black cottonwood, Populus trichocarpa (Torr. & Gray)**. *Science* 2006, **313**:1596-1604.
8. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L *et al.*: **Genome sequence of the palaeopolyploid soybean**. *Nature* 2010, **463**:178-183.
9. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B *et al.*: **The B73 maize genome: complexity, diversity, and dynamics**. *Science* 2009, **326**:1112-1115.
10. Lee JM, Sonnhammer EL: **Genomic gene clustering analysis of pathways in eukaryotes**. *Genome Res* 2003, **13**:875-882.
11. Liu X, Han B: **Evolutionary conservation of neighbouring gene pairs in plants**. *Gene* 2009, **437**:71-79.
12. Ren XY, Fiers MW, Stiekema WJ, Nap JP: **Local coexpression domains of two to four genes in the genome of Arabidopsis**. *Plant Physiol* 2005, **138**:923-934.
13. Ren XY, Stiekema WJ, Nap JP: **Local coexpression domains in the genome of rice show no microsynteny with Arabidopsis domains**. *Plant Mol Biol* 2007, **65**:205-217.
14. Williams EJ, Bowles DJ: **Coexpression of neighboring genes in the genome of Arabidopsis thaliana**. *Genome Res* 2004, **14**:1060-1067.
15. Xu Z, Kohel RJ, Song G, Cho J, Alabady M, Yu J, Koo P, Chu J, Yu S, Wilkins TA, Zhu Y, Yu JZ: **Gene-rich islands for fiber development in the cotton genome**. *Genomics* 2008, **92**:173-183.
16. Zhan S, Horrocks J, Lukens LN: **Islands of co-expressed neighbouring genes in Arabidopsis thaliana suggest higher-order chromosome domains**. *Plant J* 2006, **45**:347-357.

17. Carmel L, Koonin EV: **A Universal Nonmonotonic Relationship between Gene Compactness and Expression Levels in Multicellular Eukaryotes**. *Genome Biology and Evolution* 2009, **1**:382-390.
18. Vinogradov AE: **'Genome design' model and multicellular complexity: golden middle**. *Nucleic Acids Research* 2006, **34**:5906-5914.
19. Woody JL, Severin AJ, Bolon Y-T, Joseph B, Diers BW, Farmer AD, Weeks N, Muehlbauer GJ, Nelson RT, Grant D, Specht JE, Graham MA, Cannon SB, May GD, Vance CP, Shoemaker RC: **Gene expression patterns are correlated with genomic and genic structure in soybean**. *Genome* 2011, **54**:10-18.
20. Neale D, Wegrzyn J, Stevens K, Zimin A, Puiu D, Crepeau M, Cardeno C, Koriabine M, Holtz-Morris A, Liechty J, Martinez-Garcia P, Vasquez-Gross H, Lin B, Zieve J, Dougherty W, Fuentes-Soriano S, Wu L-S, Gilbert D, Marcais G, Roberts M, Holt C, Yandell M, Davis J, Smith K, Dean J, Lorenz W, Whetten R, Sederoff R, Wheeler N, McGuire P *et al.*: **Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies**. *Genome Biology* 2014, **15**:R59.
21. Chen ZJ, Scheffler BE, Dennis E, Triplett BA, Zhang T, Guo W, Chen X, Stelly DM, Rabinowicz PD, Town CD, Arioli T, Brubaker C, Cantrell RG, Lacape J-M, Ulloa M, Chee P, Gingle AR, Haigler CH, Percy R, Saha S, Wilkins T, Wright RJ, Van Deynze A, Zhu Y, Yu S, Abdurakhmonov I, Katageri I, Kumar PA, Mehboob-ur-Rahman, Zafar Y *et al.*: **Toward Sequencing Cotton (Gossypium) Genomes**. *Plant Physiology* 2007, **145**:1303-1310.
22. Ling HQ, Zhao S, Liu D, Wang J, Sun H, Zhang C, Fan H, Li D, Dong L, Tao Y, Gao C, Wu H, Li Y, Cui Y, Guo X, Zheng S, Wang B, Yu K, Liang Q, Yang W, Lou X, Chen J, Feng M, Jian J, Zhang X, Luo G, Jiang Y, Liu J, Wang Z, Sha Y *et al.*: **Draft genome of the wheat A-genome progenitor Triticum urartu**. *Nature* 2013, **496**:87-90.
23. Jia J, Zhao S, Kong X, Li Y, Zhao G, He W, Appels R, Pfeifer M, Tao Y, Zhang X, Jing R, Zhang C, Ma Y, Gao L, Gao C, Spannagl M, Mayer KF, Li D, Pan S, Zheng F, Hu Q, Xia X, Li J, Liang Q, Chen J, Wicker T, Gou C, Kuang H, He G, Luo Y *et al.*: **Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation**. *Nature* 2013, **496**:91-95.
24. Pfeifer M, Kugler KG, Sandve SR, Zhan B, Rudi H, Hvidsten TR, Mayer KF, Olsen OA: **Genome interplay in the grain transcriptome of hexaploid bread wheat**. *Science* 2014, **345**:1250091.
25. Gillies SA, Futardo A, Henry RJ: **Gene expression in the developing aleurone and starchy endosperm of wheat**. *Plant Biotechnol J* 2012, **10**:668-679.
26. Pellny TK, Lovegrove A, Freeman J, Tosi P, Love CG, Knox JP, Shewry PR, Mitchell RA: **Cell walls of developing wheat starchy endosperm: comparison of composition and RNA-Seq transcriptome**. *Plant Physiol* 2012, **158**:612-627.
27. Bartoš J, Paux E, Kofler R, Havránková M, Kopecký D, Suchánková P, Šafář J, Šimková H, Town CD, Lelley T, Feuillet C, Doležel J: **A first survey of the rye (*Secale cereale*) genome composition through BAC end sequencing of the short arm of chromosome 1R**. *BMC Plant Biology* 2008, **8**:95.
28. Ergen NZ, Thimmapuram J, Bohnert HJ, Budak H: **Transcriptome pathways unique to dehydration tolerant relatives of modern wheat**. *Funct Integr Genomics* 2009, **9**:377-396.
29. Choulet F, Wicker T, Rustenholz C, Paux E, Salse J, Leroy P, Schlub S, Le Paslier MC, Magdelenat G, Gonthier C, Couloux A, Budak H, Breen J, Pumphrey M, Liu S, Kong X, Jia J, Gut M, Brunel D, Anderson JA, Gill BS, Appels R, Keller B, Feuillet C: **Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces**. *Plant Cell* 2010, **22**:1686-1701.
30. Rustenholz C, Choulet F, Laugier C, Safár J, Simková H, Dolezel J, Magni F, Scalabrin S, Cattonaro F, Vautrin S, Bellec A, Bergès H, Feuillet C, Paux E, Safar J,

Simkova H, Berges H: **A 3,000-loci transcription map of chromosome 3B unravels the structural and functional features of gene islands in hexaploid wheat**. *Plant physiology* 2011, **157**:1596-1608.

31. Raats D, Frenkel Z, Krugman T, Dodek I, Sela H, Simkova H, Magni F, Cattonaro F, Vautrin S, Berges H, Wicker T, Keller B, Leroy P, Philippe R, Paux E, Dolezel J, Feuillet C, Korol A, Fahima T: **The physical map of wheat chromosome 1BS provides insights into its gene space organization and evolution**. *Genome Biology* 2013, **14**:R138.
32. The International Wheat Genome Sequencing Consortium: **A chromosome-based draft sequence of the hexaploid bread wheat genome**. *Science* 2014, **345**:1251788.
33. Brenchley R, Spannagl M, Pfeifer M, Barker GL, D'Amore R, Allen AM, McKenzie N, Kramer M, Kerhornou A, Bolser D, Kay S, Waite D, Trick M, Bancroft I, Gu Y, Huo N, Luo MC, Sehgal S, Gill B, Kianian S, Anderson O, Kersey P, Dvorak J, McCombie WR, Hall A, Mayer KF, Edwards KJ, Bevan MW, Hall N: **Analysis of the bread wheat genome using whole-genome shotgun sequencing**. *Nature* 2012, **491**:705-710.
34. Jaakson K, Zernant J, Kulm M, Hutchinson A, Tonisson N, Glavac D, Ravnik-Glavac M, Hawlina M, Meltzer MR, Caruso RC, Testa F, Maugeri A, Hoyng CB, Gouras P, Simonelli F, Lewis RA, Lupski JR, Cremers FP, Allikmets R: **Genotyping microarray (gene chip) for the ABCR (ABCA4) gene**. *Human Mutation* 2003, **22**:395 - 403.
35. Zernant J, Kulm M, Dharmaraj S, den Hollander AI, Perrault I, Preising MN, Lorenz B, Kaplan J, Cremers FP, Maumenee I, Koenekoop RK, Allikmets R: **Genotyping microarray (disease chip) for Leber congenital amaurosis: detection of modifier alleles**. *Invest Ophthalmol Vis Sci* 2005, **46**:3052 - 3059.
36. Choulet F, Alberti A, Theil S, Glover N, Barbe V, Daron J, Pingault L, Sourdille P, Couloux A, Paux E, Leroy P, Mangenot S, Guilhot N, Le Gouis J, Balfourier F, Alaux M, Jamilloux V, Poulain J, Durand C, Bellec A, Gaspin C, Safar J, Dolezel J, Rogers J, Vandepoele K, Aury J-M, Mayer K, Berges H, Quesneville H, Wincker P *et al.*: **Structural and functional partitioning of bread wheat chromosome 3B**. *Science* 2014, **345**:1249721.
37. International Rice Genome Sequencing Project: **The map-based sequence of the rice genome**. *Nature* 2005, **436**:793-800.
38. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang HB, Wang XY, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang LF, Carpita NC *et al.*: **The Sorghum bicolor genome and the diversification of grasses**. *Nature* 2009, **457**:551-556.
39. Vogel JP, Garvin DF, Mockler TC, Schmutz J, Rokhsar D, Bevan MW, Barry K, Lucas S, Harmon-Smith M, Lail K, Tice H, Grimwood J, McKenzie N, Huo NX, Gu YQ, Lazo GR, Anderson OD, You FM, Luo MC, Dvorak J, Wright J, Febrer M, Idziak D, Hasterok R, Lindquist E, Wang M, Fox SE, Priest HD, Filichkin SA, Givan SA *et al.*: **Genome sequencing and analysis of the model grass Brachypodium distachyon**. *Nature* 2010, **463**:763-768.
40. Ma L, Chen C, Liu X, Jiao Y, Su N, Li L, Wang X, Cao M, Sun N, Zhang X, Bao J, Li J, Pedersen S, Bolund L, Zhao H, Yuan L, Wong GK-S, Wang J, Deng XW, Wang J: **A microarray analysis of the rice transcriptome and its comparison to Arabidopsis**. *Genome Research* 2005, **15**:1274-1283.
41. Lu T, Lu G, Fan D, Zhu C, Li W, Zhao Q, Feng Q, Zhao Y, Guo Y, Huang X, Han B: **Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq**. *Genome Res* 2010, **20**:1238-1249.

42. Sekhon RS, Lin H, Childs KL, Hansey CN, Buell CR, de Leon N, Kaeppler SM: **Genome-wide atlas of transcription during maize development**. *Plant J* 2011, **66**:553-563.
43. Sekhon RS, Briskine R, Hirsch CN, Myers CL, Springer NM, Buell CR, de Leon N, Kaeppler SM: **Maize gene atlas developed by RNA sequencing and comparative evaluation of transcriptomes based on RNA sequencing and microarrays**. *PLoS One* 2013, **8**:e61005.
44. Libault M, Farmer A, Joshi T, Takahashi K, Langley RJ, Franklin LD, He J, Xu D, May G, Stacey G: **An integrated transcriptome atlas of the crop model Glycine max, and its use in comparative analyses in plants**. *Plant J* 2010, **63**:86-99.
45. He P, Friebe B, Gill B, Zhou J-M: **Allopolyploidy alters gene expression in the highly stable hexaploid wheat**. *Plant Molecular Biology* 2003, **52**:401-414.
46. Xu Y, Zhong L, Wu X, Fang X, Wang J: **Rapid alterations of gene expression and cytosine methylation in newly synthesized Brassica napus allopolyploids**. *Planta* 2009, **229**:471-483.
47. Zou C, Lehti-Shiu MD, Thibaud-Nissen F, Prakash T, Buell CR, Shiu S-H: **Evolutionary and Expression Signatures of Pseudogenes in Arabidopsis and Rice**. *Plant Physiology* 2009, **151**:3-15.
48. Thibaud-Nissen F, Ouyang S, Buell CR: **Identification and characterization of pseudogenes in the rice gene complement**. *BMC Genomics* 2009, **10**:317.
49. Vicient C: **Transcriptional activity of transposable elements in maize**. *BMC Genomics* 2010, **11**:601.
50. Li W, Yang W, Wang X-J: **Pseudogenes: Pseudo or Real Functional Elements?***Journal of Genetics and Genomics* 2013, **40**:171-177.
51. Wen Y-Z, Zheng L-L, Qu L-H, Ayala FJ, Lun Z-R: **Pseudogenes are not pseudo any more**. *RNA Biology* 2012, **9**:27-32.
52. Liu J, Jung C, Xu J, Wang H, Deng S, Bernad L, Arenas-Huertero C, Chua N-H: **Genome-Wide Analysis Uncovers Regulation of Long Intergenic Noncoding RNAs in Arabidopsis**. *The Plant Cell Online* 2012, **24**:4333-4345.
53. Shuai P, Liang D, Tang S, Zhang Z, Ye C-Y, Su Y, Xia X, Yin W: **Genome-wide identification and functional prediction of novel and drought-responsive lincRNAs in Populus trichocarpa**. *Journal of Experimental Botany* 2014.
54. Britto-Kido SdA, Ferreira Neto JRC, Pandolfi V, Marcelino-Guimarães FC, Nepomuceno AL, Vilela Abdelnoor R, Benko-Iseppon AM, Kido EA: **Natural Antisense Transcripts in Plants: A Review and Identification in Soybean Infected with Phakopsora pachyrhizi SuperSAGE Library**. *The Scientific World Journal* 2013, **2013**:14.
55. Zhang Y-C, Chen Y-Q: **Long noncoding RNAs: New regulators in plant development**. *Biochemical and Biophysical Research Communications* 2013, **436**:111-114.
56. Wang XJ, Gaasterland T, Chua NH: **Genome-wide prediction and identification of cis-natural antisense transcripts in Arabidopsis thaliana**. *Genome Biol* 2005, **6**:R30.
57. Coram TE, Settles ML, Chen X: **Large-scale analysis of antisense transcription in wheat using the Affymetrix GeneChip Wheat Genome Array**. *BMC Genomics* 2009, **10**:253.
58. Poole RL, Barker GLA, Werner K, Biggi GF, Coghill J, Gibbings JG, Berry S, Dunwell JM, Edwards KJ: **Analysis of Wheat Sage Tags Reveals Evidence for Widespread Antisense Transcription**. *Bmc Genomics* 2008, **9**:(10 October 2008).
59. Osato N, Yamada H, Satoh K, Ooka H, Yamamoto M, Suzuki K, Kawai J, Carninci P, Ohtomo Y, Murakami K, Matsubara K, Kikuchi S, Hayashizaki Y: **Antisense transcripts with rice full-length cDNAs**. *Genome Biol* 2003, **5**:R5.

60. Lu T, Zhu C, Lu G, Guo Y, Zhou Y, Zhang Z, Zhao Y, Li W, Lu Y, Tang W, Feng Q, Han B: **Strand-specific RNA-seq reveals widespread occurrence of novel cis-natural antisense transcripts in rice**. *BMC Genomics* 2012, **13**:721.
61. Zhang X, Lii Y, Wu Z, Polishko A, Zhang H, Chinnusamy V, Lonardi S, Zhu JK, Liu R, Jin H: **Mechanisms of small RNA generation from cis-NATs in response to environmental and developmental cues**. *Mol Plant* 2013, **6**:704-715.
62. Abranches R, Beven AF, Aragon-Alcaide L, Shaw PJ: **Transcription sites are not correlated with chromosome territories in wheat nuclei**. *J Cell Biol* 1998, **143**:5-12.
63. Baker K, Bayer M, Cook N, Dreißig S, Dhillon T, Russell J, Hedley PE, Morris J, Ramsay L, Colas I, Waugh R, Steffenson B, Milne I, Stephen G, Marshall D, Flavell AJ: **The low recombining pericentromeric region of barley restricts gene diversity and evolution but not gene expression**. *The Plant Journal* 2014:n/a-n/a.
64. Wang L, Zhao S, Gu C, Zhou Y, Zhou H, Ma J, Cheng J, Han Y: **Deep RNA-Seq uncovers the peach transcriptome landscape**. *Plant Mol Biol* 2013, **83**:365-377.
65. Werner A: **Biological functions of natural antisense transcripts**. *BMC Biology* 2013, **11**:31.
66. Nishizawa M, Okumura T, Ikeya Y, Kimura T: **Regulation of inducible gene expression by natural antisense transcripts**. *Front Biosci (Landmark Ed)* 2012, **17**:938-958.
67. Faghihi MA, Wahlestedt C: **Regulatory roles of natural antisense transcripts**. *Nat Rev Mol Cell Biol* 2009, **10**:637-643.
68. Li SW, Feng L, Niu DK: **Selection for the miniaturization of highly expressed genes**. *Biochem Biophys Res Commun* 2007, **360**:586-592.
69. Vinogradov AE: **Compactness of human housekeeping genes: selection for economy or genomic design?***Trends Genet* 2004, **20**:248-253.
70. Marquez Y, Brown JW, Simpson C, Barta A, Kalyna M: **Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis**. *Genome Res* 2012, **22**:1184-1195.
71. Walters B, Lum G, Sablok G, Min XJ: **Genome-wide landscape of alternative splicing events in Brachypodium distachyon**. *DNA Res* 2013, **20**:163-171.
72. Wang L, Cao C, Ma Q, Zeng Q, Wang H, Cheng Z, Zhu G, Qi J, Ma H, Nian H, Wang Y: **RNA-seq analyses of multiple meristems of soybean: novel and alternative transcripts, evolutionary and functional implications**. *BMC Plant Biology* 2014, **14**:169.
73. Mayer KF, Waugh R, Brown JW, Schulman A, Langridge P, Platzer M, Fincher GB, Muehlbauer GJ, Sato K, Close TJ, Wise RP, Stein N: **A physical, genetic and functional sequence assembly of the barley genome**. *Nature* 2012, **491**:711-716.
74. Syed NH, Kalyna M, Marquez Y, Barta A, Brown JW: **Alternative splicing in plants--coming of age**. *Trends Plant Sci* 2012, **17**:616-623.
75. Barbazuk WB, Fu Y, McGinnis KM: **Genome-wide analyses of alternative splicing in plants: opportunities and challenges**. *Genome Res* 2008, **18**:1381-1392.
76. Keren H, Lev-Maor G, Ast G: **Alternative splicing and evolution: diversification, exon definition and function**. *Nat Rev Genet* 2010, **11**:345-355.
77. Wang BB, Brendel V: **Genomewide comparative analysis of alternative splicing in plants**. *Proc Natl Acad Sci U S A* 2006, **103**:7175-7180.
78. Eisenberg E, Levanon EY: **Human housekeeping genes are compact**. *Trends Genet* 2003, **19**:362-365.
79. Woody JL, Shoemaker RC: **Gene expression: sizing it all up**. *Front Genet* 2011, **2**:70.
80. Urrutia AO, Hurst LD: **The signature of selection mediated by expression on human genes**. *Genome Res* 2003, **13**:2260-2264.

81. Seoighe C, Gehring C, Hurst LD: **Gametophytic selection in Arabidopsis thaliana supports the selective model of intron length reduction**. *PLoS Genet* 2005, **1**:e13.
82. Hansey CN, Vaillancourt B, Sekhon RS, de Leon N, Kaeppler SM, Buell CR: **Maize (Zea mays L.) genome diversity as revealed by RNA-sequencing**. *PLoS One* 2012, **7**:e33071.
83. Sémon M, Duret L: **Evolutionary Origin and Maintenance of Coexpressed Gene Clusters in Mammals**. *Molecular Biology and Evolution* 2006, **23**:1715-1723.
84. Spellman PT, Rubin GM: **Evidence for large domains of similarly expressed genes in the Drosophila genome**. *J Biol* 2002, **1**:5.
85. Chen WH, de Meaux J, Lercher MJ: **Co-expression of neighbouring genes in Arabidopsis: separating chromatin effects from direct interactions**. *BMC Genomics* 2010, **11**:178.
86. Parada L, McQueen P, Misteli T: **Tissue-specific spatial organization of genomes**. *Genome Biology* 2004, **5**:R44.
87. Branco MR, Pombo A: **Chromosome organization: new facts, new models**. *Trends Cell Biol* 2007, **17**:127-134.
88. Elcock LS, Bridger JM: **Exploring the relationship between interphase gene positioning, transcriptional regulation and the nuclear matrix**. *Biochemical Society Transactions* 2010, **38**:263-267.
89. Dong F, Jiang J: **Non-Rabl patterns of centromere and telomere distribution in the interphase nuclei of plant cells**. *Chromosome Research* 1998, **6**:551-558.
90. Cowan CR, Carlton PM, Cande WZ: **The Polar Arrangement of Telomeres in Interphase and Meiosis. Rabl Organization and the Bouquet**. *Plant Physiol* 2001, **125**:532-538.
91. Santos AP, Abranches R, Stoger E, Beven A, Viegas W, Shaw PJ: **The architecture of interphase chromosomes and gene positioning are altered by changes in DNA methylation and histone acetylation**. *J Cell Sci* 2002, **115**:4597-4605.
92. **FastQC** [http://www.bioinformatics.babraham.ac.uk/projects/fastqc]
93. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation**. *Nat Biotechnol* 2010, **28**:511-515.
94. **TopHat** [http://tophat.cbcb.umd.edu/]
95. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biol* 2009, **10**:R25.
96. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**:2078-2079.
97. **Cufflinks** [http://cufflinks.cbcb.umd.edu/]
98. Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J, Chen Y, Huang X, Lin W, Ye C, Tong W, Cong L, Geng J, Han Y, Li L, Li W, Hu G, Li J, Liu Z, Qi Q, Li T, Wang X *et al.*: **A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*)**. *Science* 2002, **296**:79-92.
99. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, Macmanes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, Leduc RD, Friedman N, Regev A: **De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis**. *Nat Protoc* 2013, **8**:1494-1512.
100. **changepoint-package** [http://www.inside-r.org/packages/cran/changepoint/docs/changepoint]
101. **R software** [www.r-project.org]

**Conclusions Article n°2**

En combinant les données de la pseudomolécule, avec les données RNA-Seq couvrant le développement du blé, nous avons pu acquérir de nouvelles connaissances sur les relations entre la structure des gènes, la localisation sur le chromosome, ainsi que leur régulation.

Nous avons montré que pour les 5 185 prédictions et les 3 692 NTR, le niveau de transcription est uniforme le long du chromosome, et donc pas limité aux régions distales. A partir de 3 692 NTR, nous avons mis en évidence 596 grands ARN intergéniques non codants potentiels. De plus, sur la base de données RNA-Seq orientées, nous avons détecté des transcrits anti-sens pour 12,2% des prédictions, pouvant donc être impliqués dans la régulation de l'expression des gènes.

La relation qui unie le niveau d'expression des gènes à leurs caractéristiques structurales est non monotone, et se traduit par une corrélation négative entre le niveau et l'amplitude d'expression des gènes fortement/constitutivement exprimés avec leur taille. Le pendant de cette relation étant que les gènes exprimés dans des conditions-spécifiques et/ou faiblement exprimés ont tendance à être plus grand, cela en maximisant les régions intra-géniques non codantes. Ces relations peuvent s'expliquer par la combinaison de deux modèles complémentaires et non exclusifs : la sélection pour l'économie et le dessin génomique.

Enfin, nous avons également montré que le partitionnement du chromosome basé sur la recombinaison et la densité en gènes et éléments transposables s'appliquait également aux caractéristiques structurales et fonctionnelles des gènes eux-mêmes. Ainsi les régions distales auront tendance à contenir des gènes plus petits, plus faiblement exprimés et plus spécifiques que les régions proximales. Cela se traduit également au niveau du profil d'expression de ces gènes, avec des différences marquées entre les différentes régions en termes de cluster d'expression.

Les résultats issus de ce travail apportent de nouvelles données sur l'organisation structurale et fonctionnelle du chromosome 3B. Mais ils soulèvent également de nouvelles questions concernant notamment la structure des autres chromosomes de blé ou encore les origines de ce partitionnement.

# CONCLUSIONS & PERSPECTIVES

## 1 Le chromosome 3B organisé en 3 blocs majeurs

Les travaux présentés dans cette thèse ont permis d'étudier l'espace génique du chromosome 3B de blé tendre, en se basant sur la première pseudomolécule construite pour un chromosome de blé hexaploïde. Et cela, en combinaison avec des données RNA-Seq couvrant 15 conditions de développement de la plante.

Dans un premier temps, la production d'une annotation de la pseudomolécule du chromosome 3B a permis de mettre en évidence 7 264 prédictions ordonnées le long du chromosome, classées en deux catégories : gène complet ou pseudogène. Grâce à ces résultats, nous avons confirmé que la densité de gènes n'était pas uniforme le long du chromosome, et qu'elle était croissante le long de l'axe centromère-télomère, allant de 1,3 à 27,9 gènes par Mb. Ce résultat étant l'inverse de celui observé pour les ET, qui présentent une plus forte proportion dans partie centrale du chromosome (article n°1, (Daron et al., s. d.)). De plus, nous avons montré que la proportion en gènes et pseudogènes était variable le long du chromosome avec dans trois régions (une centromérique et deux distales), une proportion de pseudogènes plus importante que la moyenne.

Grâce aux données de RNA-Seq couvrant 5 organes à trois stades de développement différents, nous avons montré que 71,4% des prédictions portées par le chromosome 3B sont exprimées dans au moins une condition. De plus, la distribution des gènes exprimés suit le gradient de densité des prédictions, avec une augmentation selon l'axe centromère-télomère. Au total, l'analyse a permis de mettre en évidence 30 232 transcrits, avec 61,4% des gènes exprimés qui ont au moins deux transcrits alternatifs. Nous avons aussi montré que la forme majoritaire d'épissage alternative était la rétention d'intron, comme observé chez le riz ou *A. thaliana*. De plus, l'analyse des profils d'expression des transcrits suggère une spécialisation des isoformes en fonction des organes ou des stades de développement. L'approche RNA-Seq nous a aussi permis de mettre en évidence 3 692 loci précédemment non annotés ou nouvelles régions transcrites (NTR). Parmi ces NTR, nous avons observé une absence de cadre de lecture et d'homologie avec des gènes codants des protéines pour 596 loci, pouvant ainsi correspondre à la catégorie des longs ARN intergéniques non codants (ou lincRNA). En plus de ces NTR, nous avons aussi mis en évidence un transcrit en orientation inverse pour 12,2% des prédictions.

Nous avons montré que, si le nombre de gènes exprimés n'est pas significativement différent d'une condition à une autre, tous les gènes ne sont pas exprimés dans les 15 tissus. En nous basant sur la position des gènes le long du chromosome, nous avons montré que

l'amplitude d'expression par gène (nombre de condition dans lequel un gène est exprimé) est négativement corrélée avec la distance des gènes par rapport au centromère et nous avons observé la même tendance pour le nombre de transcrits alternatifs exprimés par gènes. Suggérant ainsi une spécificité de l'expression des gènes situés en positions distales du chromosome. Cette hypothèse a été renforcée par des analyses d'ontologie des gènes, qui montre que les régions distales sont enrichies en gènes impliqués dans des mécanismes adaptatifs.

En nous basant sur les données de recombinaison du chromosome ainsi que sur l'annotation des ET, nous avons montré que le chromosome pouvait être séparé en cinq régions : R1, R2a, C, R2b et R3, présentant des organisations structurales et fonctionnelles très contrastées. Nous avons montré que les régions distales R1 et R3 différent de la région proximale R2a-RC-R2b, en terme de : taux de recombinaison, densité de gènes, densité d'éléments transposables, nombre de conditions exprimées par gènes, nombre de transcrits alternatif par gènes. De plus, nous avons aussi mis en évidence une relation entre la structure des gènes et le niveau d'expression, montrant ainsi que les gènes des régions R1 et R3 avaient un nombre d'exon plus faible, une taille d'introns ainsi que des transcrits plus courte que les gènes localisés dans les régions R2a, R2b et RC. Ces résultats montrent une organisation structurale et fonctionnelle du chromosome 3B jamais observée dans les génomes végétaux et suggèrent donc une évolution accélérée du génome de blé dans les régions distales de ses chromosomes, notamment au travers de la duplication et de la translocation de gènes, dits non synténiques. A partir des données de structure des gènes, du niveau ainsi que de l'amplitude d'expression, nous avons montré que deux modèles évolutifs peuvent expliquer cette organisation, et sont retrouvés combinés sur le chromosome : le modèle de la sélection pour l'économie et le modèle dessin génomique.

Ainsi, les résultats obtenus au cours de cette thèse apportent un éclairage nouveau sur l'organisation structurale et fonctionnelle de l'espace génique du génome de blé ainsi que sur son évolution. Ils ont permis la mise en évidence d'une compartimentation chromosomique jamais décrite à ce jour, participant ainsi à améliorer notre connaissance et notre compréhension de ce génome complexe. Dans la continuité de ce travail, des travaux sont désormais entrepris pour essayer d'identifier les mécanismes à l'origine de cette structuration particulière.

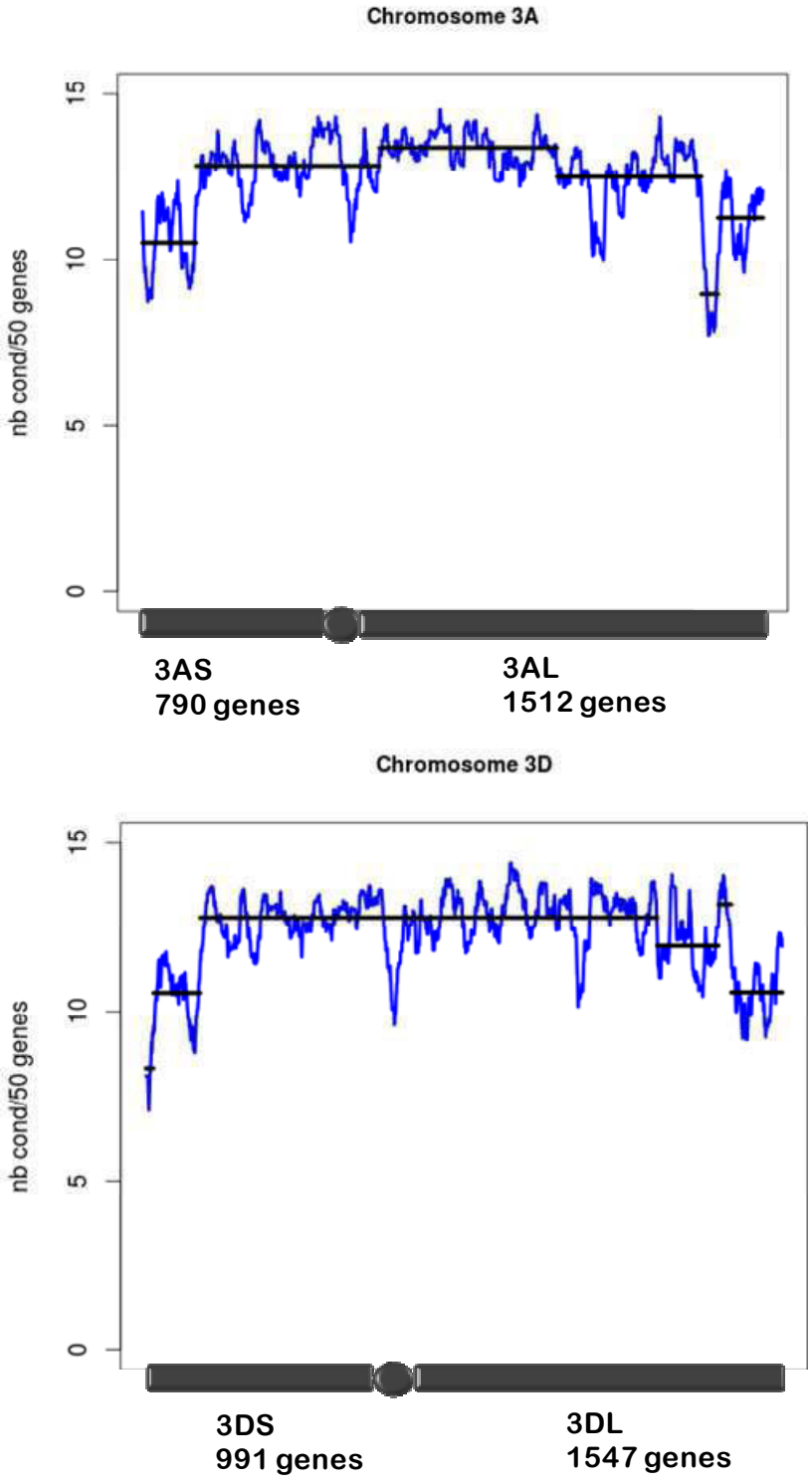



**Figure 27** : Segmentation des chromosomes 3A et 3D
Les gènes ont été ordonnés selon les données du « genome zipper » et les paramètres de segmentation sont identiques à ceux utilisés dans l'article n°2. La courbe représente l'amplitude d'expression moyenne pour 50 gènes le long de chaque chromosome.

# 2 Organisation de l'espace génique de l'ensemble des chromosomes et expression des gènes homéologues

## 2.1 Hypothèses sur l'organisation de l'espace génique des 20 autres chromosomes

L'organisation observée sur le chromosome 3B est retrouvée pour tous les chromosomes de l'orge (Article n°1). On peut donc s'attendre à retrouver un profil d'organisation en trois régions commun pour les 20 autres chromosomes de blé. Cependant, l'estimation de la taille des 20 autres chromosomes de blé varie entre 605 Mb (1D) et 928 Mb (2B). Est-ce que cette variation de taille va influencer l'organisation de l'espace génique ? Et est ce que les réarrangements chromosomiques de certains chromosomes vont contribuer à une organisation particulière ?

Afin d'apporter un début de réponse à ces questions, nous nous sommes appuyés sur l'assemblage issu du séquençage « shotgun », l'annotation (International Wheat Genome Sequencing Consortium, 2014), ainsi que sur les données RNA-Seq des 15 conditions de développement que nous avons utilisées dans les articles n°1 et n°2. Bien que les séquences soient fragmentées, notre approche a permis de montrer que les chromosomes homéologues du groupe 3 partageaient cette organisation en trois régions (Figure 27). Si nous n'avons pas retrouvé clairement cette structuration chez tous les autres chromosomes, nous avons pu néanmoins observer que pour le chromosome 4A, les différents réarrangements chromosomiques se superposent avec les régions retrouvées avec l'analyse de segmentation. Cependant, l'ordre des gènes étant basé sur le « genome zipper », les résultats obtenus sont biaisés en faveur des gènes synténiques.

Dans l'article n°2, nous avons montré que certaines régions du chromosome étaient enrichies en gènes exprimés dans un organe spécifique. En reprenant la même méthodologie pour les résultats d'expression obtenus pour les chromosomes 3A et 3D, nous n'avons pas pu retrouver cette compartimentation. Une explication possible de ces résultats reposent sur l'ordre des gènes, qui se base sur des relations de synténie, et sont en moyenne exprimés dans 11 conditions. Alors que la spécificité d'expression est apportée principalement par les gènes non synténiques. Bien que ce travail préliminaire ne puisse pas être exploité en vue d'une analyse approfondie de l'espace génique, il permet de montrer un profil global de son organisation, et pourra être implémenté en fonction de l'amélioration du « genome zipper », grâce à des marqueurs supplémentaires développés dans des projets parallèles.

## 2.2 Données sur les autres chromosomes

Grâce au projet pilot 3BSEQ, une stratégie de séquençage, d'assemblage et d'annotation d'un chromosome de blé a montré la faisabilité de séquencer un chromosome unique pour construire une séquence ainsi qu'une annotation de référence.

Le chemin est donc tracé pour les 20 autres chromosomes de blé. Ces chromosomes ont déjà été triés par bras, en utilisant des lignées aneuploïdes et ont déjà servi pour le séquençage par le méthode Illumina (2x100 pb) pour la construction d'un assemblage « shotgun » (International Wheat Genome Sequencing Consortium, 2014). La construction des cartes physiques est déjà bien engagée et il est prévu qu'elles soient terminées à la fin de l'année 2014 (« IWGSC ; http://www.wheatgenome.org »). En parallèle, le séquençage des chromosomes 1A, 3A, 4A, 7A, 1B, 4B, 6B, 7B, 3D et 7D a d'ores et déjà commencé.

L'accès aux pseudomolécules de tous les chromosomes permettra donc de faire les analyses sur une séquence très peu fragmentée, et où l'ordre des gènes ne sera pas biaisé par la conservation des gènes, à la différence de l'approche « genome zipper ». De plus, l'annotation de tout le génome suivant la même méthode permettra de comparer les résultats obtenus sans biais de méthodologie.

## 2.3 Expression des gènes homéologues

Toujours sur la base des séquences produites par l'IWGSC, nous avons cherché à exploiter au maximum ces données pour l'analyse de l'expression des gènes homéologues. Pour cela, nous avons choisi de faire une première analyse sur les chromosomes 3A, 3B et 3D. Pour les chromosomes 3A et 3D, nous nous sommes appuyé sur les données produites par l'IWGSC. L'annotation a mis en évidence 4 637 gènes et 4 546 gènes pour les chromosomes 3A et 3D respectivement. Et pour le chromosome 3B, nous avons utilisé les résultats obtenus grâce à l'analyse de la pseudomolécule.

A partir de ces données, nous avons amorcé un travail en vue d'une analyse de l'expression des gènes homéologues. En effet, 984 triplets homéologues possédant une copie unique pour chaque gène sur les trois chromosomes ont été identifiés. A partir d'une analyse de clustering hiérarchique basée sur les profils d'expression des triplets dans les 15 conditions de développement issues de l'analyse RNA-Seq, nous avons observé que 24% des triplets avaient un profil identique pour deux chromosomes, et 17% ont un profil différent pour les chromosomes. De manière intéressante, nous avons observé à partir des résultats des clusters d'expression, que le pourcentage de gènes pour la combinaison : 3A=3B≠3D (15%)

A

B

**Figure 28** : Photos de deux variété de blé tendre
A : variété Chinese Spring
B : variété Renan
(Source : Centre de Ressources Biologiques Céréales à Paille, INRA Clermont-Ferrand)

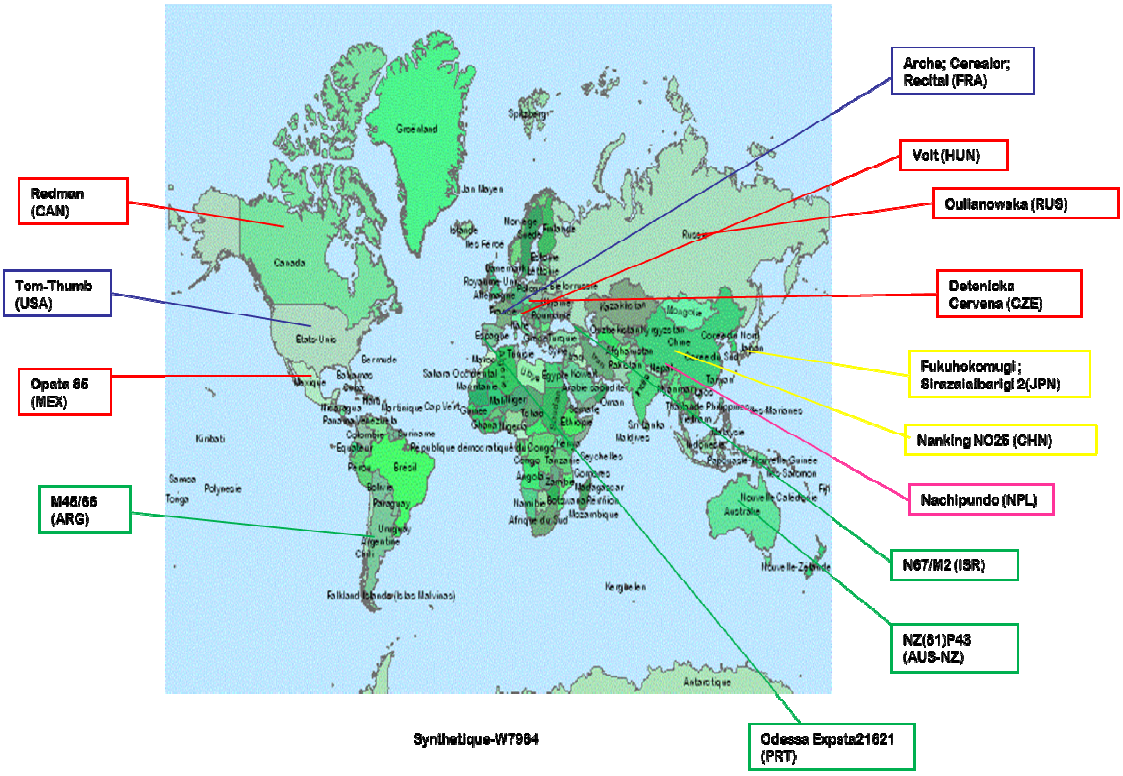
était supérieur au pourcentage des deux autres combinaisons : 3A=3D≠3B (4%) et 3D=3B≠3A (5%). L'observation d'un pourcentage plus important de gènes partageant le même cluster entre les chromosomes 3A et 3B par rapport au chromosome 3D, va dans le sens inverse de l'attendu. En effet, la fusion entre les génomes A et B étant plus ancienne, on pourrait s'attendre à ce qu'un des homéologues soit subfonctionnalisé afin de mettre en balance l'expression de ces gènes. Cependant, les résultats suggèrent que l'évolution a permis aux gènes portés par les chromosomes 3A et 3B de maintenir leur fonction initiale. Cependant, cette analyse ne porte que sur un sous set de gènes et n'est donc pas représentative de l'ensemble du génome. En effet, le nombre de gènes présents sur les chromosomes 3A et 3D est sous-estimé du fait de la fragmentation de la séquence de référence. De plus, le taux de duplications inter-chromosomiques est extrêmement élevé chez le blé (35% ; (Glover et al., s. d.)), il serait intéressant de regarder l'expression des gènes homéologues présents en plus d'une copie sur chaque chromosome.

Avec l'accès aux séquences des génomes de *T. urartu* et *Ae. tauschii*, il serait intéressant de savoir si ces mêmes gènes ont conservé leur profil d'expression. Car à la vue du fort taux de duplications intra-chromosomiques, on peut s'attendre à ce que les gènes aient évolué dans leur expression (pseudogénisation, sub ou néofonctionalisation, mutations). Ces analyses pourront aussi dévier sur la mise en évidence d'éléments *trans*-régulateurs, comme par exemple des mécanismes épigénétiques.

## 3 Organisation de l'espace génique chez d'autres variétés et à différents niveaux de ploïdie

Bien que la variété Chinese Spring (CS) soit utilisée comme référence pour les analyses génomiques du blé hexaploïde, elle n'est pas cultivée dans le monde du fait de sa faible valeur agronomique (Figure 28). Un sondage réalisé par France AgriMer fin 2012 auprès de 5 000 producteurs de blé tendre, révèle que les 5 variétés les plus cultivées en France sont dans l'ordre : Apache, Arezzo, Altigo, Expert et Pakito. L'étude indique aussi qu'il y a une tendance à la diversification des variétés cultivées. Afin de savoir si l'organisation de l'espace génique est conservée entre les variétés de blé, l'étude d'autres variétés du genre *T. aestivum*, ainsi que des variétés à des niveaux de ploïdie différent est nécessaire.

A




B

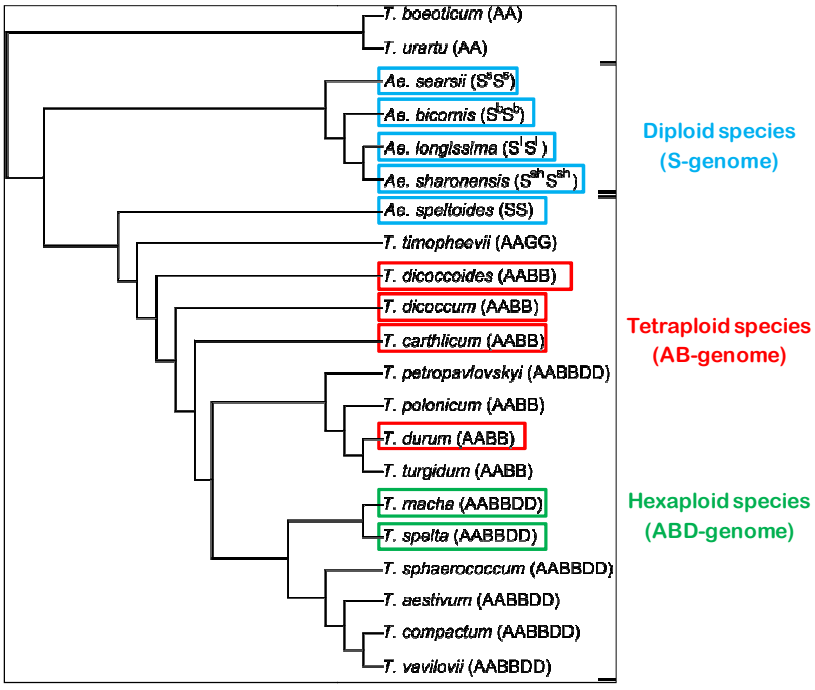



**Figure 29** : Les différentes origines géographiques des variétés de blé.
A : carte représentant les différentes localisation des variétés de blé héxaploïde.
B : arbre phylogénique des variétés de blé diploïdes, tétraploïdes, pour lesquelles le chromosome 3B a été trié et séquencé.

## 3.1 Variation du nombre de copies des gènes

Afin d'étudier les CNV chez le blé, une approche par séquençage a été initiée dans l'équipe. En effet, les chromosomes 3B ont été triés pour 44 variétés de blé : 20 *T. aestivum*, 3 *T. spelta*, 3 *T. macha*, 7 *T. durum*, 4 *T. dicoccum*, 4 *T. carthlicucum* et 3 *T. dicoccoïdes*. Les variétés *T. aestivum* étant représentatives de la diversité du blé (Figure 29). Les chromosomes ont été séquencés avec la technologie Illumina HiSeq 2000 en « paired-end » 2x100 pb, avec une couverture de 20 à 30X pour chaque chromosome. Afin de réduire la complexité due aux éléments répétés, l'analyse des CNV sera faite sur la fraction « low-copy » du chromosome 3B. Un premier alignement des lectures du séquençage des variétés hexaploïdes sur le chromosome 3B a été réalisé, et en moyenne 10% des lectures ont pu être alignés. La fraction « low-copy » correspond à 15% du chromosome 3B, on s'attend à pouvoir aligner environ 15% des lectures. Des améliorations sont donc à faire au niveau du choix de l'outil d'alignement utilisé ainsi que des paramètres.

L'étude des CNV permettra de compléter le catalogue de gènes prédits sur le chromosome 3B de la variété CS. Puis, de mettre en évidence les gènes préférentiellement affectés par les CNV du fait de leur localisation/structure/fonction. Ces résultats pourront être mis en relation avec les données déjà collectées pour ce chromosome en terme d'environnement de ET, de taux de recombinaison, de synténie, de polymorphismes (SNP) et de structure des gènes (nombre intron/exon, épissage alternatif). Compte tenu des résultats obtenus pour le chromosome 3B de la variété CS en terme d'ontologie des gènes, et de spécificité tissulaire d'expression des gènes, on peut s'attendre à ce que les régions télomériques et sub-télomériques soient les plus enrichies en CNV, car c'est aussi dans ces régions que l'on retrouve la plus forte proportion de gènes dupliqués. De plus, les gènes dont le nombre de copies est variable, sont généralement des gènes dont la fonction est en relation avec l'adaptation à l'environnement, ainsi qu'à la résistance aux maladies (Zmieńko, Samelak, Kozłowski, & Figlerowicz, 2014).

Les CNV peuvent aussi être apparentés à des marqueurs moléculaires, permettant ainsi de reconstruire la phylogénie des variétés. Chez le riz, l'analyse des CNV sur un panel de 20 variétés de riz asiatiques a montré que les variations structurales étaient spécifiques des groupes étudiés (Yu et al., 2013). Les 44 variétés de blé étant réparties dans des zones géographiques différentes couvrant les cinq continents, une étude similaire à celle menée chez le riz pourra être réalisée, afin de mettre en évidence des différences au niveau des gènes impliqués dans l'adaptation aux conditions environnementales, ainsi que des gènes sélectionnés pour leur qualité farinière. Pour cela, l'annotation des CNV pourra être mise en

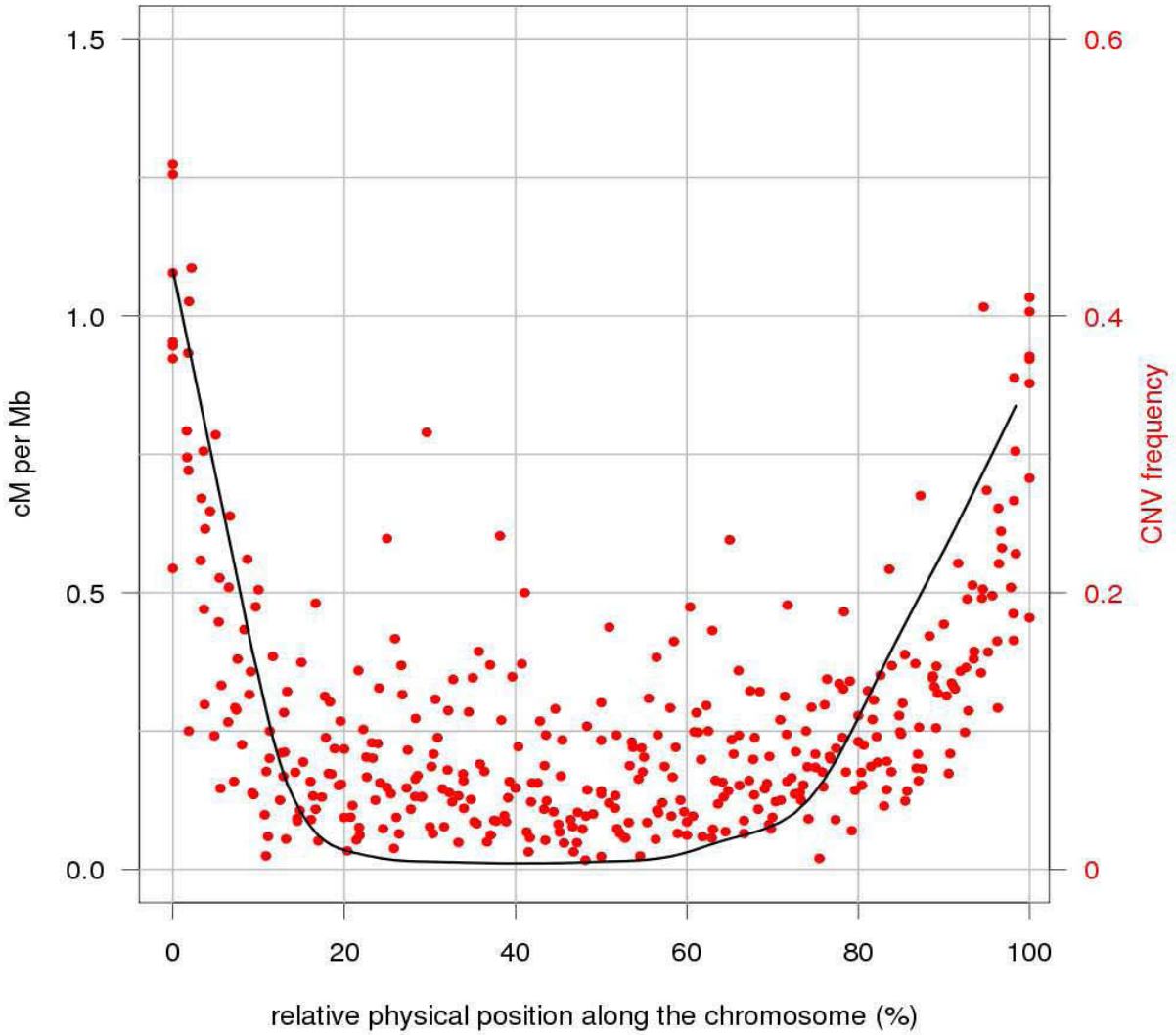



**Figure 30** : Relation entre taux de recombinaison et fréquence des CNV chez l'orge.
Les points rouges indiquent la fréquence des CNV et la courbe noire le taux de recombinaison calculé en cM par Mb. En abscisse est représenté la taille relative de tous les chromosomes en fenêtre de 10 Mb. D'après (Muñoz-Amatriaín et al. 2013).

relation avec l'annotation fonctionnelle des gènes. Ce travail pourra être aussi réalisé pour d'autres chromosomes quand les pseudomolécules auront été construites. En effet, des données de re-séquençage du génome entier ont déjà été produites et ont servit à la détection de SNP. Cependant, la gestion des paramètres d'alignement sera une étape cruciale, due fait des séquences répétés, du fort taux de duplication, ainsi que des relations d'homéologie.

L'analyse de CNV-Seq permettra aussi de mettre en évidence les CNV qui affectent directement l'expression d'un gène. Pour le génome humain, il a été montré qu'il y a une corrélation positive entre le nombre de copie d'un gène et son niveau d'expression (Haraksingh & Snyder, 2013). Chez l'orge, la distribution des CNV sur les chromosomes a montré qu'ils sont localisés dans les zones où le taux de recombinaison est élevé (Muñoz-Amatriaín et al., 2013). On s'attend donc à une organisation identique pour le chromosome 3B (Figure 30). Des insertions/délétions pourront aussi être mises en évidence, ainsi que les variations structurales qui se trouvent dans les régions régulatrices du gène et affectent son expression. Cependant, les variations de l'expression d'un gène ne peuvent pas uniquement être expliquées par les variations génétiques. D'autres mécanismes épigénétique comme la méthylation de l'ADN ou bien les changements d'état de la chromatine affectent aussi l'expression des gènes.

En parallèle, l'étude du transcriptome de ces mêmes variétés pourrait être réalisée. Ces données pourraient servir à l'analyse de la structure des chromosomes d'autres variétés de blé, et permettre ainsi de savoir si les profils des chromosomes sont spécifiques à une variété, en reliant ces données à leur spécificité phénotype, géographie, ainsi qu'à la complexité de leur génome. Ainsi que de mettre en évidence des environnements particuliers au niveau des points de segmentation, ET particuliers, enrichissement en certains motifs répétés, marques épigénétiques, profils de méthylation. Et ces données pourront être mises en relation avec les particularités des blés étudiés ainsi que leur localisation géographique.

## 3.2 Caractérisation du pan génome

L'approche CNV-Seq sur la fraction « low copy » du chromosome 3B permettra aussi de caractériser le « core genome » (séquences d'ADN qui sont présentent dans toutes les variétés étudiées) ainsi que le « dispensable genome » (segment d'ADN partiellement partagés entre les variétés ou bien variété spécifique), qui composent le pan-génome.

Ce qui pourrait permettre d'associer les gènes décrits dans le pan-génome : aux variations génétiques, ainsi qu'à la diversité phénotypique impliqués dans des caractères important d'adaptation du blé. Chez le maïs, entre les variétés B73 (2,5 Gb) et Mo17 (2,5 Gb), il a été montré que dans le pan génome, le « core genome » représente 50% du génome qui est partagé entre les deux variétés (1,67 Gb). Et que le « dispensable genome » a une taille identique dans les deux variétés (Morgante, De Paoli, & Radovic, 2007). L'étude indique aussi que le « core genome » est majoritairement composé de séquences en simple copie ainsi que de transposons. De plus, les gènes qui composent le « core genome » sont plus conservés au niveau de leur fonction que les gènes du « dispensable genome ».

A la vue de la complexité du génome de blé, ainsi que sa capacité d'adaptation à différents environnements, on peut s'attendre à ce que les loci qui composent le « core genome » soient conservés durant l'évolution et donc localisés dans la partie centrale des chromosomes (si ces chromosomes se structurent comme le chromosome 3B). Les loci du « dispensable genome » sont certainement des loci plus récents, non conservés et localisés dans les parties distales des chromosomes. Les premiers résultats vont dans ce sens. En effet, l'analyse des duplications en tandem, des insertions, des inversions ainsi que des délétions entre les chromosomes 3B de différentes variétés montrent qu'il y a un enrichissement des variations structurales dans les régions R1 et R3 du chromosome, en comparaison avec la variété CS.

## 4 ARN non codant : détection et évolution

Dans l'article n°1, grâce aux données de RNA-Seq, nous avons mis en évidence 3 692 NTR, parmi lesquels nous avons mis en évidence 596 loci peuvent être caractérisés en lincRNA.

L'accès aux séquences des pseudomolécules de l'ensemble des chromosomes permettra de caractériser l'ensemble des lincRNA. Grâce à l'analyse du chromosome 3B, on peut d'ores et déjà faire une estimation du nombre lincRNA présent dans l'ensemble du génome. Si le nombre de lincRNA est proportionnel à la taille de chaque chromosome, on s'attend à détecter pour les 15 conditions de développement analysées, environ 32 200 lincRNA. Avec le nombre croissant de données de RNA-Seq, notamment dans des conditions de développement particulières, ce nombre va certainement augmenter. Ces données permettront de savoir si les lincRNA présentent un comportement identique à celui des gènes codant des protéines en termes de relation d'homéologie, de ratio, de taux de duplication inter-chromosomique, ainsi qu'au niveau de leur profil d'expression.

Chez les tétrapodes, il semble que seulement 3% des lncRNA soient conservés chez les différentes espèces. Cette même analyse a révélé que 81% des 13 533 des familles de lncRNA sont conservés entre les mammifères, et 19% ont une origine de plus de $90x10^6$ ans (Necsulea et al., 2014). Dans l'article n°2, nous avons montré que les lincRNA étaient moins conservés que les gènes codant des protéines. Une analyse plus en profondeur permettra de dater l'origine des lincRNA.

Chez les eucaryotes, différents mécanismes de formation ont déjà été mis à évidence : le mouvement des ET, les réarrangements chromosomiques, les duplications par rétrotransposition, les duplications en tandem (Ponting, Oliver, & Reik, 2009). Mais est ce que ces mécanismes sont retrouvés chez le blé? Le lien avec l'annotation des ET (Daron et al., s. d.) ainsi que la caractérisation des évènements de duplications (Glover et al., s. d.) permettra de répondre à cette question.

Une catégorie des petits ARN non codants n'a cependant pas encore était annotée pour le chromosome 3B : les miRNA, qui sont impliqués, entre autre dans les mécanismes de régulation de l'expression des gènes, ainsi que de méthylation de l'ADN. En plus de leur détection *in silico,* des librairies de petits ARN sont déjà disponibles pour le stade feuille de la variété CS (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM803792), et pourront être utilisées pour l'annotation des miRNA. Cependant, d'autres librairies pourraient être construites à partir d'autres conditions de développement « classiques » de la plante, mais aussi pour des conditions plus spécifiques, comme par exemple des conditions de stress a biotiques. Ce travail sera réalisé en collaboration avec Christine Gaspin (INRA, Toulouse) et Hikmet Budak (Sanbanci U, Turquie), et débutera à la fin de cette année.

Le paysage transcriptionnel des smallRNA pourra ainsi venir s'ajouter aux données d'expression, de méthylation et des marques épigénétiques. Ce qui permettra de fournir des réponses quant à la régulation de l'expression des gènes et plus particulièrement celle des gènes homéologues (régulation *cis* vs. *trans*). Une étude sur le poisson zèbre, a montré que les miRNA étaient aussi préférentiellement associés aux parties 3' des lncRNA (Jalali, Bhartiya, Lalwani, Sivasubbu, & Scaria, 2013). L'analyse de ces données chez le blé pourrait permettre de savoir si cette association est identique dans un génome polyploïde.

Enfin, la création de base de données pour l'ensemble des catégories d'ARN non codants sera pourra aussi être utilisée par les outils d'annotation des génomes des Triticeae.

# 5 Co expression des gènes : mécanismes mis en jeu

## 5.1 Détections des promoteurs

Dans l'article n°2, nous avons montré que chez le blé à l'échelle du chromosome, les gènes regroupés en îlots sont co-exprimés. Cependant, l'approche RNA-Seq ne nous permet pas de savoir si les gènes co-exprimés partagent des séquences régulatrices de leur expression, ainsi que les sites d'initiation de la transcription (TSS, « Transcription Start Site »), ni de mettre en évidence les promoteurs spécifique de l'expression (PSE) qui contiennent des sites spécifique à la fixation des facteurs de transcription, comme les régions activatrices, qui sont essentielles au contrôle de l'expression tissue/condition spécifique. Pour cela, des librairies de CAGE-Seq pourraient être construites pour des conditions identiques à celles utilisées pour le RNA-Seq. La technologie de CAGE ou « Cap Analysis of Gene Expression » a été introduite en 2003 pour détecter les sites de début de la transcription. Cette technologie se base sur la capture des ARN possédant une coiffe en 5' (les ARNt et ARNr ne sont donc pas captés), à la différence du RNA-Seq qui sélectionne les ARN sur leur queue polyA en 3'(Shiraki et al., 2003).

Par cette méthode, il serait possible de définir à la base près les TSS ainsi que les séquences promotrices de tous les ARN possédant une coiffe, qu'ils soient codant ou non codant, et ainsi relier ces données avec le niveau d'expression des gènes. Chez les génomes de l'homme et de la souris, l'analyse du CAGE a permis de déterminer que les promoteurs caractérisés par une TATA-box ont un site de début de transcription unique, alors que les promoteurs reliés à un îlot CpG ont plusieurs site de début de transcription répartis sur une grande région.

Le CAGE est aussi une bonne approche pour étudier les promoteurs bidirectionnels, qui peuvent être co-régulés en partageant des sites de fixation aux facteurs de transcription. L'étude de la dynamique des promoteurs en comparant la distribution du niveau d'expression en plus des TSS dans la région en amont du gènes peut permettre de mettre en évidence les promoteurs qui sont activés ou inactivés pour différentes conditions de développement analysées (de Hoon & Hayashizaki, 2008). Une étude publiée en 2014 par Kawaji a cherché à comparer le CAGE avec un séquençage de deuxième (Illumina) ou troisième (Heliscope) génération au RNA-Seq (Kawaji et al., 2014). Les conclusions de cette étude montrent qu'il y a une bonne reproductibilité entre les différentes plateformes au niveau de l'expression des gènes. La combinaison des deux techniques CAGE-Seq et RNA-Seq permet d'avoir une approche extrêmement résolutive sur les variations non connues de la structure des

transcrits. Ces deux techniques sont complémentaires pour mettre en évidence et affiner la complexité de la structure des transcrits ainsi que de leurs unités régulatrices.

## 5.2 Les marques épigénétiques impliquées dans la régulation de l'expression

Le terme « épigénétique » est utilisé pour décrire les études sur l'héritage de l'information qui ne peut pas être expliquée par les variations de la séquence ADN (Haig, 2004). Sous le terme épigénétique, on peut différencier deux domaines : la mémoire d'un état de l'expression d'un gène pendant une condition de développement ou dans une condition environnementale (héritage mitotique, hétérochromatine facultative) et la mémoire trans-générationnelle d'un état de l'expression d'un gène (héritage méïotique, hétérochromatine constitutive).

Les études épigénomiques à la différence des études de génomiques (organisation de la séquence ADN dans le génome), portent sur les ajouts chimiques qui affectent la structure de la chromatine, et comment ces modifications affectent l'organisation de l'information contenue dans un génome (Mirouze & Vitte, 2014). Cependant, il faut différencier le méthylome qui se restreint à la méthylation de l'ADN, de l'épigénome qui combine les modifications d'histones ainsi que la méthylation des cytosines (Mirouze & Vitte, 2014).

L'activité transcriptionnelle des gènes pour les eucaryotes est déterminée par l'action combinée des facteurs de transcription et des protéines modifiant la chromatine. Presque toutes les cellules d'un organisme partagent le même génome, cependant il en résulte différents phénotypes et fonctions. Les types individuels de cellules qui sont caractérisées par des profiles d'expression différents, sont générés durant le développement et sont maintenu de manière stable. L'état chromatinien, c'est à dire la compaction de l'ADN avec les protéines histone et non-histone, a un effet sur l'expression d'un gène et est considéré comme contributeur à l'établissement et au maintien de l'identité des cellules, et les transitions au niveau du développement sont accompagnées de changements de l'état de la chromatine. Et dans la vision d'aujourd'hui de la régulation des gènes, il faut tenir compte du fait que l'ADN eucaryote est compactée autour d'un noyau de protéines : les histones, pour former le nucléosome (Kornberg, 1974), qui est formé d'histones en octamères, contenant deux copies des histones : H2A, H2B, H3 et H4 autour desquels 147 pb d'ADN est enroulé.

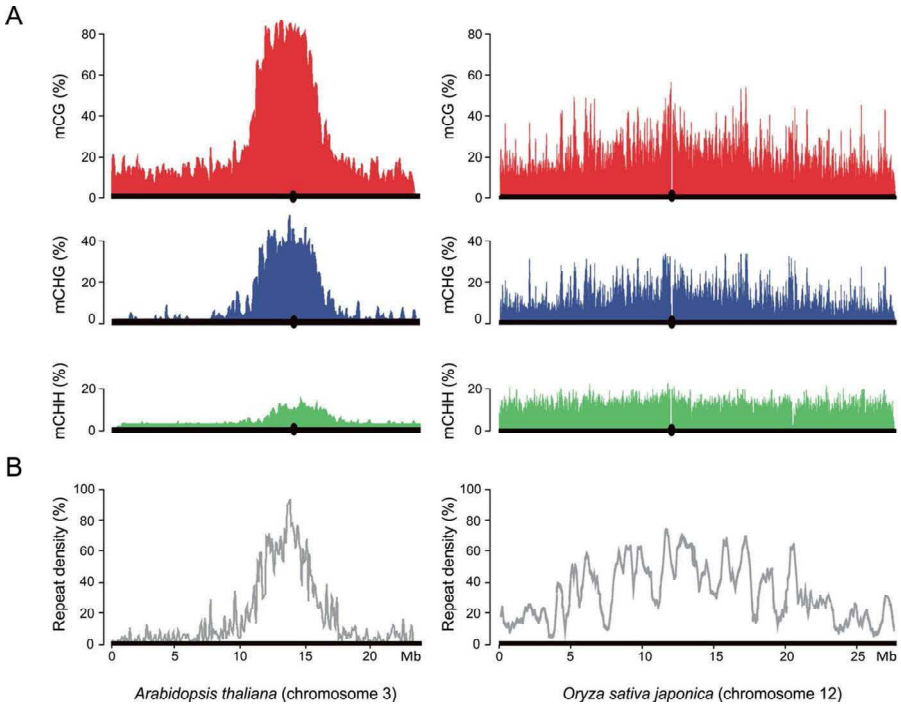



**Figure 31** : Paysage de la méthylation de l'ADN et de la densité des TE pour les chromosomes de riz et d'*Arabidopsis*.
(A) Distribution de la méthylation des cytosines à l'échelle du chromosome sur un fenêtre glissante de 100 kb pour Arabidopsis et 50 kb pour le riz, rouge : mCG, bleu : mCHG, vert : mCHH.
(B) Distribution de la du pourcentage de la densité des éléments répétés sur une échelle glissante de 100 kb. D'après (Mirouze et Vitte 2014).

La chromatine représente une barrière physique pour la transcription et les facteurs capables de modifier la structure et la composition de cette chromatine peuvent ainsi activer ou réprimer l'expression d'un gène. Il a été montré qu'il y a une corrélation négative entre l'expression des gènes et la méthylation des cytosines (mC), qui peut réprimer la transcription en altérant la structure de la chromatine. Cependant l'altération de la chromatine n'affecte pas la séquence d'ADN primaire, on parle d'épigénétique si le changement au niveau du nouveau chromosome formé n'est pas hérité après que le stimulus l'ayant provoqué est éliminé. Les mécanismes qui régulent la structure de la chromatine inclus : mC, les facteurs qui affectent le composition ainsi que la position du nucléosome et les modifications post-transcriptionnelles (MPTs) des histones (Lauria & Rossi, 2011). Les récentes avancées avec les nouvelles technologies de séquençage ont permis de faire une avancée considérable dans la compréhension du paysage épigénétique chez les plantes.

### 5.2.1 Etude du méthylome

Des études sur la méthylation de la cytosine chez Arabidopsis, le riz, le maïs, le soja ou bien la tomate ont déjà été publiées (X. Li et al., 2012; Regulski et al., 2013; Reinders & Paszkowski, 2009; Schmitz et al., 2013). Les résultats montrent que les régions enrichies en ET ainsi qu'en éléments répétés sont plus fortement méthylées. De plus, les résultats obtenus à partir des différents méthylomes montrent que les marques de méthylation sont retrouvées sur les motifs : CG, CHG, CHH (H=A ;T ;C), alors que le corps des gènes les marques de méthylation sont retrouvées uniquement sur les motifs GC (Mirouze & Vitte, 2014). Les résultats obtenus dans l'article n°1 montrent que chez le blé tendre, les régions centromériques et péri centromériques sont enrichies en ET, on s'attend donc à ce que les marques de méthylation suivent les profils obtenus chez les méthylomes des Angiospermes déjà analysés. Cependant, Mirouze et Vitte indiquent dans leur revue que si les profils de méthylation semblent conservés dans leurs grandes lignes entre les différents génomes étudiés, chaque espèce montre un profil particulier. Car si la méthylation semble concentrée au niveau du centromère sur les chromosomes d'Arabidopsis, elle beaucoup plus étalée sur les chromosomes du riz, mais corrèle toujours à la densité des TE (Figure 31) (Mirouze & Vitte, 2014).

### 5.2.2 Caractérisation du paysage épigénétique du chromosome 3B

Le paysage épigénétique du blé n'a pas encore été décrit à l'heure actuelle, du fait de l'absence de séquence de référence. Afin de réduire la redondance du génome ainsi que les coûts de séquençage, une approche par capture de séquence a été initiée dans l'équipe. La capture des séquences a été construite afin de cibler uniquement sur le chromosome 3B, les régions codantes et les lncRNA, avec les séquences situées 2 kb en amont et en aval de chaque loci. Les différentes marques étudiées sont des marques activatrices de l'expression des gènes (H3K4me2, H3K36me3 et H3K4me3), ainsi que des marques de répression de la chromatine (H3K27me3, 5meC). L'organe pour lequel les marques épigénétiques vont être étudiées est la feuille. Le stade de développement choisi a été le stade trois feuille, afin de pouvoir mettre en relation les résultats de ChIP-Seq et de RNA-Seq précédemment décrits dans les articles n°1 et n°2. La stratégie de séquençage choisie pour le ChIP-Seq est une approche Illumina, HiSeq2500 en 2x100 pb avec une profondeur de 60X.

Les marques épigénétiques sont dépendantes d'un tissu ou bien d'une condition, donc pour le stade de développement utilisé pour l'analyse du paysage épigénomique on ne s'attend pas à avoir des marques de condensation/décondensation de la chromatine sur tout le chromosome car il n'y a qu'un seul tissu étudié. Dans article n°2, nous avons montré que les régions du chromosome portent des gènes spécifiquement exprimés dans un tissu ou à un temps de développement donné. Dans cet article, nous avons aussi montré que l'expression des gènes dans les feuilles était associée à un cluster d'expression particulier situé dans la région R1 du chromosome. En ce que concerne l'analyse des marques d'activation de la transcription, on s'attend donc à ce quelles soient plus présentes dans cette région du chromosome, alors que les marques de répression seraient retrouvées en plus grande proportion dans les régions R2 et R3.

Les mises au points du protocole de capture des marques épigénétiques se révèlent déjà concluant de pars l'analyse des premiers résultats. Avec l'accès aux séquences des gènes annotés sur les autres chromosomes, un kit de capture pourra être développé cibler l'ensemble des marques présentent sur le génome du blé au niveau des séquences transcrites. De plus, l'analyse pourra être réalisée sur un plus grand nombre de tissus et de conditions de développement.

En conclusion, les travaux de ma thèse ont permis de confirmer les hypothèses ainsi que d'obtenir de nouveaux résultats quant à l'organisation de l'espace génique ainsi que la régulation des gènes, sur la base de la première séquence complète d'un chromosome de blé hexaploïde. Ce travail va aussi servir de base pour l'analyse des mécanismes de régulation des l'expression des gènes, ainsi que pour l'annotation de transcrits non codants. Ces travaux ont aussi été utilisé dans l'analyse de la fraction répétée du chromosome 3B, ainsi que pour celle des gènes dupliqués. En parallèle, nous avons aussi utilisé les données produites en RNA-Seq pour l'analyse de l'expression des copies du gènes MET1.

# BIBLIOGRAPHIE

Adams, K. L., & Wendel, J. F. (2005). Polyploidy and genome evolution in plants. *Current Opinion in Plant Biology*, *8*(2), 135‑141. doi:10.1016/j.pbi.2005.01.001

Akhunov, E. D., Akhunova, A. R., & Dvořák, J. (2005). BAC libraries of Triticum urartu, Aegilops speltoides and Ae. tauschii, the diploid ancestors of polyploid wheat. *Theoretical and Applied Genetics*, *111*(8), 1617‑1622. doi:10.1007/s00122-005-0093-1

Akhunov, E. D., Goodyear, A. W., Geng, S., Qi, L.-L., Echalier, B., Gill, B. S., … Dvorak. (2003). The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. *Genome research*, *13*(5), 753‑63. doi:10.1101/gr.808603

Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, *408*(6814), 796‑815. doi:10.1038/35048692

Bainbridge, M. N., Warren, R. L., Hirst, M., Romanuik, T., Zeng, T., Go, A., … Jones, S. J. M. (2006). Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics*, *7*, 246. doi:10.1186/1471-2164-7-246

Balakirev, E. S., & Ayala, F. J. (2003). Pseudogenes: are they « junk » or functional DNA? *Annual Review of Genetics*, *37*, 123‑151. doi:10.1146/annurev.genet.37.040103.103949

Barakat, A., Carels, N., & Bernardi, G. (1997). The distribution of genes in the genomes of Gramineae. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(13), 6857‑6861.

Barbazuk, W. B., Fu, Y., & McGinnis, K. M. (2008). Genome-wide analyses of alternative splicing in plants: opportunities and challenges. *Genome Research*, *18*(9), 1381‑1392. doi:10.1101/gr.053678.106

Benne, R., Van den Burg, J., Brakenhoff, J. P., Sloof, P., Van Boom, J. H., & Tromp, M. C. (1986). Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell*, *46*(6), 819‑826.

Bennetzen, J. L. (2002). Mechanisms and rates of genome expansion and contraction in flowering plants. *Genetica*, *115*(1), 29‑36. doi:10.1023/A:1016015913350

Bennetzen, J. L., & Wang, H. (2014). The Contributions of Transposable Elements to the Structure, Function, and Evolution of Plant Genomes. *Annual Review of Plant Biology*. doi:10.1146/annurev-arplant-050213-035811

Bento, M., Gustafson, J. P., Viegas, W., & Silva, M. (2011). Size matters in Triticeae polyploids: larger genomes have higher remodeling. *Genome / National Research Council Canada = Génome / Conseil National de Recherches Canada*, *54*(3), 175‑183. doi:10.1139/G10-107

Berget, S. M. (1995). Exon recognition in vertebrate splicing. *The Journal of Biological Chemistry*, *270*(6), 2411‑2414.

Berget, S. M., Moore, C., & Sharp, P. A. (1977). Spliced segments at the 5’ terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(8), 3171‑3175.

Blumenthal, T. (2005). Trans-splicing and operons. *WormBook*. doi:10.1895/wormbook.1.5.1

Breen, J., Wicker, T., Shatalina, M., Frenkel, Z., Bertin, I., Philippe, R., … Keller, B. (2013). A physical map of the short arm of wheat chromosome 1A. *PloS One*, *8*(11), e80272. doi:10.1371/journal.pone.0080272

Brenchley, R., Spannagl, M., Pfeifer, M., Barker, G. L. A., D’Amore, R., Allen, A. M., … Hall, N. (2012). Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, *491*(7426), 705‑710. doi:10.1038/nature11650

Bullard, J. H., Purdom, E., Hansen, K. D., & Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, *11*(1), 94. doi:10.1186/1471-2105-11-94

Bustamante, C. D., Nielsen, R., & Hartl, D. L. (2002). A maximum likelihood method for analyzing pseudogene evolution: implications for silent site evolution in humans and rodents. *Molecular Biology and Evolution*, *19*(1), 110‑117.

Chain, P. S. G., Grafham, D. V., Fulton, R. S., FitzGerald, M. G., Hostetler, J., Muzny, D., … Detter, J. C. (2009). Genome Project Standards in a New Era of Sequencing. *Science*, *326*(5950), 236‑237. doi:10.1126/science.1180614

Chantret, N., Salse, J., Sabot, F., Rahman, S., Bellec, A., Laubin, B., … Chalhoub, B. (2005). Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (Triticum and Aegilops). *The Plant Cell*, *17*(4), 1033‑1045. doi:10.1105/tpc.104.029181

Charmet, G. (2011). Wheat domestication: lessons for the future. *Comptes Rendus Biologies*, *334*(3), 212‑220. doi:10.1016/j.crvi.2010.12.013

Choulet, F., Alberti, A., Theil, S., Glover, N., Barbe, V., Daron, J., … Feuillet, C. (2014). Structural and functional partitioning of bread wheat chromosome 3B. *Science (New York, N.Y.)*, *345*(6194), 1249721. doi:10.1126/science.1249721

Choulet, F., Wicker, T., Rustenholz, C., Paux, E., Salse, J., Leroy, P., … Feuillet, C. (2010). Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. *The Plant cell*, *22*(6), 1686‑701. doi:10.1105/tpc.110.074187

Church, D. M., Goodstadt, L., Hillier, L. W., Zody, M. C., Goldstein, S., She, X., … Mouse Genome Sequencing Consortium. (2009). Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biology*, *7*(5), e1000112. doi:10.1371/journal.pbio.1000112

Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nature Reviews. Genetics*, *6*(11), 836‑846. doi:10.1038/nrg1711

Covello, P. S., & Gray, M. W. (1989). RNA editing in plant mitochondria. *Nature*, *341*(6243), 662‑666. doi:10.1038/341662a0

D'Hont, A., Denoeud, F., Aury, J.-M., Baurens, F.-C., Carreel, F., Garsmeur, O., … Wincker, P. (2012). The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature*, *488*(7410), 213‑7. doi:10.1038/nature11241

Daron, J., Glover, N., Pingault, L., Theil, S., Jamilloux, V., Paux, E., … Choulet, F. (s. d.). Organization and Evolution of Transposable Elements along the Wheat Chromosome 3B.

*DAWGPAWS - http://dawgpaws.sourceforge.net/*. (s. d.). Consulté à l'adresse http://dawgpaws.sourceforge.net/

De Hoon, M., & Hayashizaki, Y. (2008). Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference. *BioTechniques*, *44*(5), 627‑628, 630, 632. doi:10.2144/000112802

Del Fabbro, C., Scalabrin, S., Morgante, M., & Giorgi, F. M. (2013). An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PloS One*, *8*(12), e85024. doi:10.1371/journal.pone.0085024

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., … Guigo, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression, 1775‑1789. doi:10.1101/gr.132159.111

Devos, K. M., Brown, J. K. M., & Bennetzen, J. L. (2002). Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. *Genome Research*, *12*(7), 1075‑1079. doi:10.1101/gr.132102

Devos, K. M., Ma, J., Pontaroli, A. C., Pratt, L. H., & Bennetzen, J. L. (2005). Analysis and mapping of randomly chosen bacterial artificial chromosome clones from hexaploid bread wheat. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(52), 19243‑19248. doi:10.1073/pnas.0509473102

Djebali, S., Davis, C. a, Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., … Gingeras, T. R. (2012). Landscape of transcription in human cells. *Nature*, *489*(7414), 101‑8. doi:10.1038/nature11233

Eilam, T., Anikster, Y., Millet, E., Manisterski, J., Sagi-Assif, O., & Feldman, M. (2007). Genome size and genome evolution in diploid Triticeae species. *Genome / National Research Council Canada = Génome / Conseil National de Recherches Canada*, *50*(11), 1029‑1037. doi:10.1139/g07-083

Ewing, B., & Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, *8*(3), 186‑194.

*FAOSTAT*. (2014). *Food and Agriculture Organization of the United Nations*. Consulté à l'adresse http://faostat.fao.org/

*FASTX-Toolkit*. (s. d.). Consulté 12 juin 2014, à l'adresse http://hannonlab.cshl.edu/fastx_toolkit/index.html

Fatica, A., & Bozzoni, I. (2014). Long non-coding RNAs: new players in cell differentiation and development. *Nature Reviews. Genetics*, *15*(1), 7‑21. doi:10.1038/nrg3606

Feldman, M., & Levy, A. A. (2012). Genome evolution due to allopolyploidization in wheat. *Genetics*, *192*(3), 763‑774. doi:10.1534/genetics.112.146316

Feuillet, C., Leach, J. E., Rogers, J., Schnable, P. S., & Eversole, K. (2011). Crop genome sequencing: lessons and rationales. *Trends in plant science*, *16*(2), 77‑88. doi:10.1016/j.tplants.2010.10.005

Filichkin, S. a, Priest, H. D., Givan, S. a, Shen, R., Bryant, D. W., Fox, S. E., … Mockler, T. C. (2010). Genome-wide mapping of alternative splicing in Arabidopsis thaliana. *Genome Research*, *20*(1), 45‑58. doi:10.1101/gr.093302.109.2008

Francis, W. R., Christianson, L. M., Kiko, R., Powers, M. L., Shaner, N. C., & D Haddock, S. H. (2013). A comparison across non-model animals suggests an optimal sequencing depth for de novo transcriptome assembly. *BMC Genomics*, *14*, 167. doi:10.1186/1471-2164-14-167

Frankish, A., Mudge, J. M., Thomas, M., & Harrow, J. (2012). The importance of identifying alternative splicing in vertebrate genome annotation. *Database: the journal of biological databases and curation*, *2012*, bas014. doi:10.1093/database/bas014

Frederickson, R. M. (2002). Fluidigm. Biochips get indoor plumbing. *Chemistry & Biology*, *9*(11), 1161‑1162.

Garber, M., Grabherr, M. G., Guttman, M., & Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods*, *8*(6), 469‑477. doi:10.1038/nmeth.1613

Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbel, J. O., … Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. *Genome Research*, *17*(6), 669‑681. doi:10.1101/gr.6339607

Gill, B. S., Appels, R., Botha-Oberholster, A.-M., Buell, C. R., Bennetzen, J. L., Chalhoub, B., … Sasaki, T. (2004). A workshop report on wheat genome sequencing: International Genome Research on Wheat Consortium. *Genetics*, *168*(2), 1087‑96. doi:10.1534/genetics.104.034769

Gill, K. S., Gill, B. S., & Endo, T. R. (1993). A chromosome region-specific mapping strategy reveals gene-rich telomeric ends in wheat. *Chromosoma*, *102*(6), 374‑381. doi:10.1007/BF00360401

Giorgi, F. M., Fabbro, C. D., & Licausi, F. (2013). Comparative study of RNA-seq- and Microarray-derived coexpression networks in Arabidopsis thaliana. *Bioinformatics (Oxford, England)*, *2*, 1‑8. doi:10.1093/bioinformatics/btt053

Glover, N., Daron, J., Pingault, L., Vandepoele, K., Paux, E., Choulet, F., & Feuillet, C. (s. d.). Single gene duplications played a major role in the recent evolution of the hexaploid wheat genome.

Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., … Jaffe, D. B. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences*, *108*(4), 1513‑1518. doi:10.1073/pnas.1017351108

Goff, S. a, Ricke, D., Lan, T.-H., Presting, G., Wang, R., Dunn, M., … Briggs, S. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science (New York, N.Y.)*, *296*(5565), 92‑100. doi:10.1126/science.1068275

Gonçalves, I., Duret, L., & Mouchiroud, D. (2000). Nature and structure of human genes that generate retropseudogenes. *Genome Research*, *10*(5), 672‑678.

Gottlieb, A., Müller, H.-G., Massa, A. N., Wanjugi, H., Deal, K. R., You, F. M., … Dvorak, J. (2013). Insular organization of gene space in grass genomes. *PloS One*, *8*(1), e54101. doi:10.1371/journal.pone.0054101

Greilhuber, J., Borsch, T., Müller, K., Worberg, A., Porembski, S., & Barthlott, W. (2006). Smallest angiosperm genomes found in lentibulariaceae, with chromosomes of bacterial size. *Plant Biology (Stuttgart, Germany)*, *8*(6), 770‑777. doi:10.1055/s-2006-924101

Griffith, M., Griffith, O. L., Mwenifumbo, J., Goya, R., Morrissy, A. S., Morin, R. D., … Marra, M. A. (2010). Alternative expression analysis by RNA sequencing. *Nature Methods*, *7*(10), 843‑847. doi:10.1038/nmeth.1503

Griffiths, S., Sharp, R., Foote, T. N., Bertin, I., Wanous, M., Reader, S., … Moore, G. (2006). Molecular characterization of Ph1 as a major chromosome pairing locus in polyploid wheat. *Nature*, *439*(7077), 749‑752. doi:10.1038/nature04434

Guo, S., Zheng, Y., Joung, J.-G., Liu, S., Zhang, Z., Crasta, O. R., … Fei, Z. (2010). Transcriptome sequencing and comparative analysis of cucumber flowers with different sex types. *BMC Genomics*, *11*, 384. doi:10.1186/1471-2164-11-384

Guo, W., Cai, C., Wang, C., Zhao, L., Wang, L., & Zhang, T. (2008). A preliminary analysis of genome structure and composition in Gossypium hirsutum. *BMC Genomics*, *9*, 314. doi:10.1186/1471-2164-9-314

Guo, X., Zhang, Z., Gerstein, M. B., & Zheng, D. (2009). Small RNAs originated from pseudogenes: cis- or trans-acting? *PLoS Computational Biology*, *5*(7), e1000449. doi:10.1371/journal.pcbi.1000449

Guo, Y., Sheng, Q., Li, J., Ye, F., Samuels, D. C., & Shyr, Y. (2013). Large Scale Comparison of Gene Expression Levels by Microarrays and RNAseq Using TCGA Data. *PLoS ONE*, *8*(8), e71462. doi:10.1371/journal.pone.0071462

Gurtan, A. M., & Sharp, P. A. (2013). The role of miRNAs in regulating gene expression networks. *Journal of Molecular Biology*, *425*(19), 3582‑3600. doi:10.1016/j.jmb.2013.03.007

Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., … Regev, A. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology*, *28*(5), 503‑510. doi:10.1038/nbt.1633

Haig, D. (2004). The (dual) origin of epigenetics. *Cold Spring Harbor Symposia on Quantitative Biology*, *69*, 67‑70. doi:10.1101/sqb.2004.69.67

Hansey, C. N., Vaillancourt, B., Sekhon, R. S., de Leon, N., Kaeppler, S. M., & Buell, C. R. (2012). Maize (Zea mays L.) genome diversity as revealed by RNA-sequencing. *PloS One*, *7*(3), e33071. doi:10.1371/journal.pone.0033071

Haraksingh, R. R., & Snyder, M. P. (2013). Impacts of variation in the human genome on gene regulation. *Journal of Molecular Biology*, *425*(21), 3970‑3977. doi:10.1016/j.jmb.2013.07.015

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., … Guigo, R. (2012). GENCODE: The reference human genome annotation for The ENCODE Project, 1760‑1774. doi:10.1101/gr.135350.111

Hatem, A., Bozdağ, D., Toland, A. E., & Çatalyürek, Ü. V. (2013). Benchmarking short sequence mapping tools. *BMC Bioinformatics*, *14*, 184. doi:10.1186/1471-2105-14-184

Henson, J., Tischler, G., & Ning, Z. (2012). Next-generation sequencing and large genome assemblies. *Pharmacogenomics*, *13*(8), 901‑915. doi:10.2217/pgs.12.72

Heo, J. B., Lee, Y.-S., & Sung, S. (2013). Epigenetic regulation by long noncoding RNAs in plants. *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology*. doi:10.1007/s10577-013-9392-6

Hernandez, P., Martis, M., Dorado, G., Pfeifer, M., Gálvez, S., Schaaf, S., … Mayer, K. F. X. (2012). Next-generation sequencing and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content. *The Plant Journal: For Cell and Molecular Biology*, *69*(3), 377‑386. doi:10.1111/j.1365-313X.2011.04808.x

Huang, S., Li, R., Zhang, Z., Li, L., Gu, X., Fan, W., … Li, S. (2009). The genome of the cucumber, Cucumis sativus L. *Nature genetics*, *41*(12), 1275‑81. doi:10.1038/ng.475

Huang, S., Sirikhachornkit, A., Su, X., Faris, J., Gill, B., Haselkorn, R., & Gornicki, P. (2002). Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the Triticum/Aegilops complex and the evolutionary history of polyploid wheat. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(12), 8133‑8138. doi:10.1073/pnas.072223799

Iida, K., & Go, M. (2006). Survey of conserved alternative splicing events of mRNAs encoding SR proteins in land plants. *Molecular Biology and Evolution*, *23*(5), 1085‑1094. doi:10.1093/molbev/msj118

International Barley Genome Sequencing Consortium, Mayer, K. F. X., Waugh, R., Brown, J. W. S., Schulman, A., Langridge, P., … Stein, N. (2012). A physical, genetic and functional sequence assembly of the barley genome. *Nature*, *491*(7426), 711‑716. doi:10.1038/nature11543

International Brachypodium Initiative. (2010). Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature*, *463*(7282), 763‑768. doi:10.1038/nature08747

International Rice Genome Sequencing Project. (2005). The map-based sequence of the rice genome. *Nature*, *436*(7052), 793‑800. doi:10.1038/nature03895

International, T., & Initiative, B. (2010). Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature*, *463*(7282), 763‑8. doi:10.1038/nature08747

International Wheat Genome Sequencing Consortium. (2014). A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science (New York, N.Y.)*, *345*(6194), 1251788. doi:10.1126/science.1251788

Isshiki, M., Morino, K., Nakajima, M., Okagaki, R. J., Wessler, S. R., Izawa, T., & Shimamoto, K. (1998). A naturally occurring functional allele of the rice waxy locus has a GT to TT mutation

at the 5' splice site of the first intron. *The Plant Journal: For Cell and Molecular Biology*, *15*(1), 133‑138.

Isshiki, M., Tsumoto, A., & Shimamoto, K. (2006). The serine/arginine-rich protein family in rice plays important roles in constitutive and alternative splicing of pre-mRNA. *The Plant Cell*, *18*(1), 146‑158. doi:10.1105/tpc.105.037069

Iwata, H., & Gotoh, O. (2011). Comparative analysis of information contents relevant to recognition of introns in many species. *BMC Genomics*, *12*, 45. doi:10.1186/1471-2164-12-45

*IWGSC ; http://www.wheatgenome.org*. (s. d.). Consulté 8 juillet 2014, à l'adresse http://www.wheatgenome.org/

Jacq, C., Miller, J. R., & Brownlee, G. G. (1977). A pseudogene structure in 5S DNA of Xenopus laevis. *Cell*, *12*(1), 109‑120.

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., … Wincker, P. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, *449*(7161), 463‑7. doi:10.1038/nature06148

Jalali, S., Bhartiya, D., Lalwani, M. K., Sivasubbu, S., & Scaria, V. (2013). Systematic transcriptome wide analysis of lncRNA-miRNA interactions. *PloS One*, *8*(2), e53823. doi:10.1371/journal.pone.0053823

Jia, J., Zhao, S., Kong, X., Li, Y., Zhao, G., He, W., … Wang, J. (2013). Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature*, *496*(7443), 91‑95. doi:10.1038/nature12028

Kageyama, Y., Kondo, T., & Hashimoto, Y. (2011). Coding vs non-coding: Translatability of short ORFs found in putative non-coding transcripts. *Biochimie*, *93*(11), 1981‑1986. doi:10.1016/j.biochi.2011.06.024

Kandouz, M., Bier, A., Carystinos, G. D., Alaoui-Jamali, M. A., & Batist, G. (2004). Connexin43 pseudogene is expressed in tumor cells and inhibits growth. *Oncogene*, *23*(27), 4763‑4770. doi:10.1038/sj.onc.1207506

Kasschau, K. D., Fahlgren, N., Chapman, E. J., Sullivan, C. M., Cumbie, J. S., Givan, S. A., & Carrington, J. C. (2007). Genome-wide profiling and analysis of Arabidopsis siRNAs. *PLoS Biology*, *5*(3), e57. doi:10.1371/journal.pbio.0050057

Katz, Y., Wang, E. T., Airoldi, E. M., & Burge, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, *7*(12), 1009‑1015. doi:10.1038/nmeth.1528

Kawaji, H., Lizio, M., Itoh, M., Kanamori-Katayama, M., Kaiho, A., Nishiyori-Sueki, H., … FANTOM Consortium. (2014). Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing. *Genome Research*, *24*(4), 708‑717. doi:10.1101/gr.156232.113

Keller, B., & Feuillet, C. (2000). Colinearity and gene density in grass genomes. *Trends in Plant Science*, *5*(6), 246‑251.

Kelly, L. J., & Leitch, I. J. (2011). Exploring giant plant genomes with next-generation sequencing technology. *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology*, *19*(7), 939‑953. doi:10.1007/s10577-011-9246-z

Keren, H., Lev-Maor, G., & Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. *Nature reviews. Genetics*, *11*(5), 345‑55. doi:10.1038/nrg2776

Kidwell, M. G. (2002). Transposable elements and the evolution of genome size in eukaryotes. *Genetica*, *115*(1), 49‑63. doi:10.1023/A:1016072014259

Kogenaru, S., Qing, Y., Guo, Y., & Wang, N. (2012). RNA-seq and microarray complement each other in transcriptome profiling. *BMC Genomics*, *13*, 629. doi:10.1186/1471-2164-13-629

Kornberg, R. D. (1974). Chromatin structure: a repeating unit of histones and DNA. *Science (New York, N.Y.)*, *184*(4139), 868‑871.

Korneev, S. A., Park, J. H., & O'Shea, M. (1999). Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *19*(18), 7711‑7720.

Kubaláková, M., Kovárová, P., Suchánková, P., Číhalíková, J., Bartos, J., Lucretti, S., … Dolezel, J. (2005). Chromosome sorting in tetraploid wheat and its potential for genome analysis. *Genetics*, *170*(2), 823‑829. doi:10.1534/genetics.104.039180

Lander, E. S., & Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, *2*(3), 231‑239.

Lauria, M., & Rossi, V. (2011). Epigenetic control of gene regulation in plants. *Biochimica et Biophysica Acta*, *1809*(8), 369‑378. doi:10.1016/j.bbagrm.2011.03.002

Lee, T.-H., Tang, H., Wang, X., & Paterson, A. H. (2012). PGDD: a database of gene and genome duplication in plants. *Nucleic Acids Research*, gks1104. doi:10.1093/nar/gks1104

Leitch, A. R., & Leitch, I. J. (2008). Genomic plasticity and the diversity of polyploid plants. *Science (New York, N.Y.)*, *320*(5875), 481‑483. doi:10.1126/science.1153585

Leitch, A. R., & Leitch, I. J. (2012). Ecological and genetic factors linked to contrasting genome dynamics in seed plants: *Tansley review*. *New Phytologist*, *194*(3), 629‑646. doi:10.1111/j.1469-8137.2012.04105.x

Leitch, I. J., Soltis, D. E., Soltis, P. S., & Bennett, M. D. (2005). Evolution of DNA Amounts Across Land Plants (Embryophyta). *Annals of Botany*, *95*(1), 207‑217. doi:10.1093/aob/mci014

Leroy, P., Guilhot, N., Sakai, H., Bernard, A., Choulet, F., Theil, S., … Feuillet, C. (2012). TriAnnot: A Versatile and High Performance Pipeline for the Automated Annotation of Plant Genomes. *Frontiers in Plant Science*, *3*, 5. doi:10.3389/fpls.2012.00005

Li, L., Eichten, S. R., Shimizu, R., Petsch, K., Yeh, C.-T., Wu, W., … Muehlbauer, G. J. (2014). Genome-wide discovery and characterization of maize long non-coding RNAs. *Genome Biology*, *15*(2), R40. doi:10.1186/gb-2014-15-2-r40

Li, W., Yang, W., & Wang, X.-J. (2013). Pseudogenes: pseudo or real functional elements? *Journal of Genetics and Genomics = Yi Chuan Xue Bao*, *40*(4), 171‑177. doi:10.1016/j.jgg.2013.03.003

Li, X., Zhu, J., Hu, F., Ge, S., Ye, M., Xiang, H., … Wang, W. (2012). Single-base resolution maps of cultivated and wild rice methylomes and regulatory roles of DNA methylation in plant gene expression. *BMC Genomics*, *13*, 300. doi:10.1186/1471-2164-13-300

Linardopoulou, E. V., Williams, E. M., Fan, Y., Friedman, C., Young, J. M., & Trask, B. J. (2005). Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature*, *437*(7055), 94‑100. doi:10.1038/nature04029

Ling, H.-Q., Zhao, S., Liu, D., Wang, J., Sun, H., Zhang, C., … Wang, J. (2013). Draft genome of the wheat A-genome progenitor Triticum urartu. *Nature*, *496*(7443), 87‑90. doi:10.1038/nature11997

Liu, B., Yuan, J., Yiu, S.-M., Li, Z., Xie, Y., Chen, Y., … Luo, R. (2012). COPE: an accurate k-mer-based pair-end reads connection tool to facilitate genome assembly. *Bioinformatics (Oxford, England)*, *28*(22), 2870‑2874. doi:10.1093/bioinformatics/bts563

Liu, J., Jung, C., Xu, J., Wang, H., Deng, S., Bernad, L., … Chua, N.-H. (2012). Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. *The Plant Cell*, *24*(11), 4333‑4345. doi:10.1105/tpc.112.102855

Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D., & Darnell, J. (2000). *Molecular Cell Biology*. Text. Consulté 17 février 2014, à l'adresse http://www.ncbi.nlm.nih.gov/books/NBK21475/

Lohse, M., Bolger, A. M., Nagel, A., Fernie, A. R., Lunn, J. E., Stitt, M., & Usadel, B. (2012). RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research*, *40*(Web Server issue), W622‑627. doi:10.1093/nar/gks540

Lucas, S. J., Akpınar, B. A., Kantar, M., Weinstein, Z., Aydınoğlu, F., Safář, J., … Budak, H. (2013). Physical mapping integrated with syntenic analysis to characterize the gene space of the long arm of wheat chromosome 1A. *PloS One*, *8*(4), e59542. doi:10.1371/journal.pone.0059542

Macmanes, M. D. (2014). On the optimal trimming of high-throughput mRNA sequence data. *Frontiers in Genetics*, *5*, 13. doi:10.3389/fgene.2014.00013

Madlung, A. (2013). Polyploidy and its effect on evolutionary success: old questions revisited with new tools. *Heredity*, *110*(2), 99‑104. doi:10.1038/hdy.2012.79

Magadum, S., Banerjee, U., Murugan, P., Gangapur, D., & Ravikesavan, R. (2013). Gene duplication as a major force in evolution. *Journal of Genetics*, *92*(1), 155‑161.

Malone, J. H., & Oliver, B. (2011). Microarrays, deep sequencing and the true measure of the transcriptome. *BMC biology*, *9*, 34‑34. doi:10.1186/1741-7007-9-34

Marquez, Y., Brown, J. W. S., Simpson, C., Barta, A., & Kalyna, M. (2012). Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome research*, *22*(6), 1184‑1195. doi:10.1101/gr.134106.111

Martínez-Gómez, P., Crisosto, C. H., Bonghi, C., & Rubio, M. (2011). New approaches to Prunus transcriptome analysis. *Genetica*, *139*(6), 755‑69. doi:10.1007/s10709-011-9580-2

Martinez-Perez, E., Shaw, P., & Moore, G. (2001). The Ph1 locus is needed to ensure specific somatic and meiotic centromere association. *Nature*, *411*(6834), 204‑207. doi:10.1038/35075597

Massa, A. N., Wanjugi, H., Deal, K. R., O'Brien, K., You, F. M., Maiti, R., … Devos, K. M. (2011). Gene Space Dynamics During the Evolution of Aegilops tauschii, Brachypodium distachyon, Oryza sativa, and Sorghum bicolor Genomes. *Molecular Biology and Evolution*, *28*(9), 2537‑2547. doi:10.1093/molbev/msr080

Merchant, S. S., Prochnik, S. E., Vallon, O., Harris, E. H., Karpowicz, S. J., Witman, G. B., … Grossman, A. R. (2007). The Chlamydomonas Genome Reveals the Evolution of Key Animal and Plant Functions. *Science*, *318*(5848), 245‑250. doi:10.1126/science.1143609

Michael, T. P., & Jackson, S. (2013). *The First 50 Plant Genomes*. Consulté 27 mars 2014, à l'adresse https://www.crops.org/publications/tpg/articles/6/2/plantgenome2013.03.0001in

Ming, R., Hou, S., Feng, Y., Yu, Q., Dionne-Laporte, A., Saw, J. H., … Alam, M. (2008). The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). *Nature*, *452*(7190), 991‑6. doi:10.1038/nature06856

Mirouze, M., & Vitte, C. (2014). Transposable elements, a treasure trove to decipher epigenetic variation: insights from Arabidopsis and crop epigenomes. *Journal of Experimental Botany*. doi:10.1093/jxb/eru120

Moore, G. (2002). Meiosis in allopolyploids -- the importance of « Teflon » chromosomes. *Trends in Genetics: TIG*, *18*(9), 456‑463.

Moore, G., Devos, K. M., Wang, Z., & Gale, M. D. (1995). Cereal genome evolution. Grasses, line up and form a circle. *Current Biology: CB*, *5*(7), 737‑739.

Morgan, T. ., Sturtevant, A. ., Muller, H. ., & Bridges, C. . (1915). The Mechanism of Mendelian heredity. *Holt Rinehart & Winston, New York*. Consulté à l'adresse https://archive.org/stream/mechanismofmende00morgiala#page/20/mode/2up

Morgante, M., De Paoli, E., & Radovic, S. (2007). Transposable elements and the plant pan-genomes. *Current Opinion in Plant Biology*, *10*(2), 149‑155. doi:10.1016/j.pbi.2007.02.001

Mortazavi, A., Williams, B. A., Mccue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq, *5*(7), 1‑8. doi:10.1038/NMETH.1226

Mouse Genome Sequencing Consortium, Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., … Lander, E. S. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, *420*(6915), 520‑562. doi:10.1038/nature01262

Muñoz-Amatriaín, M., Eichten, S. R., Wicker, T., Richmond, T. A., Mascher, M., Steuernagel, B., … Stein, N. (2013). Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. *Genome Biology*, *14*(6), R58. doi:10.1186/gb-2013-14-6-r58

Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., … Kaessmann, H. (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*, *505*(7485), 635‑640. doi:10.1038/nature12943

Numa, H., & Itoh, T. (2014). MEGANTE: a web-based system for integrated plant genome annotation. *Plant & Cell Physiology*, *55*(1), e2. doi:10.1093/pcp/pct157

Otto, S. P. (2007). The evolutionary consequences of polyploidy. *Cell*, *131*(3), 452‑462. doi:10.1016/j.cell.2007.10.022

Otto, S. P., & Whitton, J. (2000). Polyploid incidence and evolution. *Annual Review of Genetics*, *34*, 401‑437. doi:10.1146/annurev.genet.34.1.401

Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nature reviews. Genetics*, *12*(2), 87‑98. doi:10.1038/nrg2934

P. S. Baenziger, Russell, W. K., Graef, G. L., & Campbell, B. T. (2006). Improving Lives: 50 Years of Crop Breeding, Genetics, and Cytology (C-1). *Crop Science*, *46*(5), 2230‑2244.

Pagel, M., & Johnstone, R. A. (1992). Variation across Species in the Size of the Nuclear Genome Supports the Junk-DNA Explanation for the C-Value Paradox. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *249*(1325), 119‑124. doi:10.1098/rspb.1992.0093

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., … Rokhsar, D. S. (2009a). The Sorghum bicolor genome and the diversification of grasses. *Nature*, *457*(7229), 551‑556. doi:10.1038/nature07723

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., … Rokhsar, D. S. (2009b). The Sorghum bicolor genome and the diversification of grasses. *Nature*, *457*(7229), 551‑6. doi:10.1038/nature07723

Paux, E., Sourdille, P., Salse, J., Saintenac, C., Choulet, F., Leroy, P., … Feuillet, C. (2008). A physical map of the 1-gigabase bread wheat chromosome 3B. *Science (New York, N.Y.)*, *322*(5898), 101‑4. doi:10.1126/science.1161847

Pei, B., Sisu, C., Frankish, A., Howald, C., Habegger, L., Mu, X. J., … Gerstein, M. B. (2012). The GENCODE pseudogene resource. *Genome biology*, *13*(9), R51‑R51. doi:10.1186/gb-2012-13-9-r51

Pellicer, J., Fay, M. F., & Leitch, I. J. (2010). The largest eukaryotic genome of them all? *Botanical Journal of the Linnean Society*, *164*(1), 10‑15. doi:10.1111/j.1095-8339.2010.01072.x

Ponting, C. P., Oliver, P. L., & Reik, W. (2009). Evolution and Functions of Long Noncoding RNAs. *Cell*, *136*(4), 629‑641. doi:10.1016/j.cell.2009.02.006

Rabinowicz, P. D., Citek, R., Budiman, M. A., Nunberg, A., Bedell, J. A., Lakey, N., … Martienssen, R. A. (2005). Differential methylation of genes and repeats in land plants. *Genome Research*, *15*(10), 1431‑1440. doi:10.1101/gr.4100405

Reddy, A. S. N. (2007). Alternative splicing of pre-messenger RNAs in plants in the genomic era. *Annual review of plant biology*, *58*, 267‑294. doi:10.1146/annurev.arplant.58.032806.103754

Regulski, M., Lu, Z., Kendall, J., Donoghue, M. T. A., Reinders, J., Llaca, V., … Martienssen, R. A. (2013). The maize methylome influences mRNA splice sites and reveals widespread paramutation-like switches guided by small RNA. *Genome Research*, *23*(10), 1651‑1662. doi:10.1101/gr.153510.112

Reinders, J., & Paszkowski, J. (2009). Unlocking the Arabidopsis epigenome. *Epigenetics: Official Journal of the DNA Methylation Society*, *4*(8), 557‑563.

Rensing, S. A. (2014). Gene duplication as a driver of plant morphogenetic evolution. *Current Opinion in Plant Biology*, *17*, 43‑48. doi:10.1016/j.pbi.2013.11.002

Rice, J. C., & Garcia, S. M. (2011). Fisheries, food security, climate change, and biodiversity: characteristics of the sector and perspectives on emerging issues. *ICES Journal of Marine Science: Journal Du Conseil*, fsr041. doi:10.1093/icesjms/fsr041

Rizzon, C., Ponger, L., & Gaut, B. S. (2006). Striking Similarities in the Genomic Distribution of Tandemly Arrayed Genes in Arabidopsis and Rice. *PLoS Comput Biol*, *2*(9), e115. doi:10.1371/journal.pcbi.0020115

Roulin, A., Auer, P. L., Libault, M., Schlueter, J., Farmer, A., May, G., … Jackson, S. A. (2012). The fate of duplicated genes in a polyploid plant genome. *The Plant Journal: For Cell and Molecular Biology*. doi:10.1111/tpj.12026

Roy, B., Haupt, L. M., & Griffiths, L. R. (2013). Review: Alternative Splicing (AS) of Genes As An Approach for Generating Protein Complexity. *Current Genomics*, *14*(3), 182‑194. doi:10.2174/1389202911314030004

Rustenholz, C., Choulet, F., Laugier, C., Safár, J., Simková, H., Dolezel, J., … Paux, E. (2011). A 3,000-loci transcription map of chromosome 3B unravels the structural and functional features of gene islands in hexaploid wheat. *Plant Physiology*, *157*(4), 1596‑1608. doi:10.1104/pp.111.183921

Rustenholz, C., Hedley, P. E., Morris, J., Choulet, F., Feuillet, C., Waugh, R., & Paux, E. (2010). *Specific patterns of gene space organisation revealed in wheat by using the combination of barley and wheat genomic resources.* (Vol. 11). Consulté à l'adresse http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3019236&tool=pmcentrez&rendertype=abstract

Sakata, K., Nagamura, Y., Numa, H., Antonio, B. A., Nagasaki, H., Idonuma, A., … Higo, K. (2002). RiceGAAS: an automated annotation system and database for rice genome sequence. *Nucleic Acids Research*, *30*(1), 98‑102.

Sakharkar, M. K., Chow, V. T. K., & Kangueane, P. (2004). Distributions of exons and introns in the human genome. *In Silico Biology*, *4*(4), 387‑393.

Schatz, M. C., Witkowski, J., & McCombie, W. R. (2012). Current challenges in de novo plant genome sequencing and assembly. *Genome Biology*, *13*(4), 243. doi:10.1186/gb4015

Schmieder, R., & Edwards, R. (2011). Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PloS One*, *6*(3), e17288. doi:10.1371/journal.pone.0017288

Schmitz, R. J., He, Y., Valdés-López, O., Khan, S. M., Joshi, T., Urich, M. A., … Ecker, J. R. (2013). Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. *Genome Research*, *23*(10), 1663‑1674. doi:10.1101/gr.152538.112

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., … Jackson, S. a. (2010). Genome sequence of the palaeopolyploid soybean. *Nature*, *463*(7278), 178‑83. doi:10.1038/nature08670

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., … Wilson, R. K. (2009). The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science*, *326*(5956), 1112‑1115. doi:10.1126/science.1178534

Scott Jackson, Barbara Hass Jacobus, & Janice Pagel. (2004). The Gene Space of the Soybean Genome. In *Legume Crop Genomics* (Vol. 1-0). AOCS Publishing. Consulté à l'adresse http://www.crcnetbase.com/doi/abs/10.1201/9781439822265.ch11

Sehgal, S. K., Li, W., Rabinowicz, P. D., Chan, A., Simková, H., Doležel, J., & Gill, B. S. (2012). Chromosome arm-specific BAC end sequences permit comparative analysis of homoeologous chromosomes and genomes of polyploid wheat. *BMC Plant Biology*, *12*, 64. doi:10.1186/1471-2229-12-64

Shewry, P. R. (2009). Wheat. *Journal of experimental botany*, *60*(6), 1537‑53. doi:10.1093/jxb/erp058

Shi, H., Xiong, L., Stevenson, B., Lu, T., & Zhu, J.-K. (2002). The Arabidopsis salt overly sensitive 4 mutants uncover a critical role for vitamin B6 in plant salt tolerance. *The Plant Cell*, *14*(3), 575‑588.

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., … Hayashizaki, Y. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(26), 15776‑15781. doi:10.1073/pnas.2136655100

Smeds, L., & Künstner, A. (2011). ConDeTri--a content dependent read trimmer for Illumina data. *PloS One*, *6*(10), e26314. doi:10.1371/journal.pone.0026314

Soergel, D. A. W., Lareau, L. F., & Brenner, S. E. (2006). *Regulation of Gene Expression by Coupling of Alternative Splicing and NMD*. *http://www.landesbioscience.com/curie/chapter/2833/*. Consulté 20 mars 2014, à l'adresse http://www.landesbioscience.com/curie/chapter/2833/

Soltis, D. E., Albert, V. A., Leebens-Mack, J., Bell, C. D., Paterson, A. H., Zheng, C., … Soltis, P. S. (2009). Polyploidy and angiosperm diversification. *American Journal of Botany*, *96*(1), 336‑348. doi:10.3732/ajb.0800079

Soltis, D. E., Soltis, P. S., Bennett, M. D., & Leitch, I. J. (2003). Evolution of genome size in the angiosperms. *American Journal of Botany*, *90*(11), 1596‑1603. doi:10.3732/ajb.90.11.1596

Soneson, C., & Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, *14*, 91. doi:10.1186/1471-2105-14-91

Spannagl, M., Martis, M. M., Pfeifer, M., Nussbaumer, T., & Mayer, K. F. (2013). Analysing complex Triticeae genomes - concepts and strategies. *Plant Methods*, *9*(1), 35. doi:10.1186/1746-4811-9-35

Stebbins, G. L. (1971). *Chromosomal evolution in higher plants*. Edward Arnold.

Stuart, K. (1991). RNA Editing in Trypanosomatid Mitochondria. *Annual Review of Microbiology*, *45*(1), 327‑344. doi:10.1146/annurev.mi.45.100191.001551

Syed, N. H., Kalyna, M., Marquez, Y., Barta, A., & Brown, J. W. S. (2012). Alternative splicing in plants - coming of age. *Trends in Plant Science*, *17*(10), 1‑8. doi:10.1016/j.tplants.2012.06.001

Taft, R. J., Pheasant, M., & Mattick, J. S. (2007). The relationship between non-protein-coding DNA and eukaryotic complexity. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, *29*(3), 288‑299. doi:10.1002/bies.20544

Takenaka, M., Zehrmann, A., Verbitskiy, D., Härtel, B., & Brennicke, A. (2013). RNA editing in plants and its evolution. *Annual Review of Genetics*, *47*, 335‑352. doi:10.1146/annurev-genet-111212-133519

Tenaillon, M. I., Hollister, J. D., & Gaut, B. S. (2010). A triptych of the evolution of plant transposable elements. *Trends in Plant Science*, *15*(8), 471‑478. doi:10.1016/j.tplants.2010.05.003

The Tomato Genome Consortium. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, *485*(7400), 635‑641. doi:10.1038/nature11119

This week in Nature - Editorials. (2014). Wheat lag. *Nature*, *507*(7493), 399‑400. doi:10.1038/507399b

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., … Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, *28*(5), 511‑515. doi:10.1038/nbt.1621

Tuskan, G. a, Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., … Rokhsar, D. (2006). The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science (New York, N.Y.)*, *313*(5793), 1596‑604. doi:10.1126/science.1128691

Ulitsky, I., & Bartel, D. P. (2013). lincRNAs: genomics, evolution, and mechanisms. *Cell*, *154*(1), 26‑46. doi:10.1016/j.cell.2013.06.020

Varshney, R. K., Hoisington, D. A., & Tyagi, A. K. (2006). Advances in cereal genomics and applications in crop breeding. *Trends in Biotechnology*, *24*(11), 490‑499. doi:10.1016/j.tibtech.2006.08.006

Vega, J. M., & Feldman, M. (1998). Effect of the pairing gene Ph1 on centromere misdivision in common wheat. *Genetics*, *148*(3), 1285‑1294.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., … Zhu, X. (2001). The Sequence of the Human Genome. *Science*, *291*(5507), 1304‑1351. doi:10.1126/science.1058040

Vitulo, N., Albiero, A., Forcato, C., Campagna, D., Dal Pero, F., Bagnaresi, P., … Stanca, A. M. (2011). First survey of the wheat chromosome 5A composition through a next generation sequencing approach. *PloS One*, *6*(10), e26421. doi:10.1371/journal.pone.0026421

Walters, B., Lum, G., Sablok, G., & Min, X. J. (2013). Genome-wide landscape of alternative splicing events in Brachypodium distachyon. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, *20*(2), 163‑171. doi:10.1093/dnares/dss041

Wang, B.-B., & Brendel, V. (2006). Genomewide comparative analysis of alternative splicing in plants. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(18), 7175‑7180. doi:10.1073/pnas.0602039103

Wang, K. C., & Chang, H. Y. (2011). Molecular mechanisms of long noncoding RNAs. *Molecular Cell*, *43*(6), 904‑914. doi:10.1016/j.molcel.2011.08.018

Wang, L., Zhao, S., Gu, C., Zhou, Y., Zhou, H., Ma, J., … Han, Y. (2013). Deep RNA-Seq uncovers the peach transcriptome landscape. *Plant Molecular Biology*, *83*(4-5), 365‑377. doi:10.1007/s11103-013-0093-5

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57‑63. doi:10.1038/nrg2484

Weiss-Schneeweiss, H., Emadzade, K., Jang, T.-S., & Schneeweiss, G. M. (2013). Evolutionary Consequences, Constraints and Potential of Polyploidy in Plants. *Cytogenetic and genome research*, *140*(0). doi:10.1159/000351727

Wendel, J. F., Cronn, R. C., Alvarez, I., Liu, B., Small, R. L., & Senchina, D. S. (2002). Intron Size and Genome Size in Plants. *Molecular Biology and Evolution*, *19*(12), 2346‑2352.

Wessler, S. R. (2006). Transposable elements and the evolution of eukaryotic genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(47), 17600‑1. doi:10.1073/pnas.0607612103

Whitford, R., Fleury, D., Reif, J. C., Garcia, M., Okada, T., Korzun, V., & Langridge, P. (2013). Hybrid breeding in wheat: technologies to improve hybrid wheat seed production. *Journal of Experimental Botany*, *64*(18), 5411‑5428. doi:10.1093/jxb/ert333

Xu, X., Pan, S., Cheng, S., Zhang, B., Mu, D., Ni, P., … Visser, R. G. F. (2011). Genome sequence and analysis of the tuber crop potato. *Nature*, *475*(7355), 189‑95. doi:10.1038/nature10158

Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews. Genetics*, *13*(5), 329‑342. doi:10.1038/nrg3174

Yoshimura, K., Yabuta, Y., Ishikawa, T., & Shigeoka, S. (2002). Identification of a cis element for tissue-specific alternative splicing of chloroplast ascorbate peroxidase pre-mRNA in higher plants. *The Journal of Biological Chemistry*, *277*(43), 40623‑40632. doi:10.1074/jbc.M201531200

Yu, P., Wang, C.-H., Xu, Q., Feng, Y., Yuan, X.-P., Yu, H.-Y., … Wei, X.-H. (2013). Genome-wide copy number variations in Oryza sativa L. *BMC Genomics*, *14*(1), 649. doi:10.1186/1471-2164-14-649

Zenoni, S., Ferrarini, A., Giacomelli, E., Xumerle, L., Fasoli, M., Malerba, G., … Delledonne, M. (2010). Characterization of transcriptional complexity during berry development in Vitis vinifera using RNA-Seq. *Plant physiology*, *152*(4), 1787‑95. doi:10.1104/pp.109.149716

Zhan, S., Horrocks, J., & Lukens, L. N. (2006). Islands of co-expressed neighbouring genes in Arabidopsis thaliana suggest higher-order chromosome domains. *The Plant Journal: For Cell and Molecular Biology*, *45*(3), 347‑357. doi:10.1111/j.1365-313X.2005.02619.x

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., & Liu, X. (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PloS One*, *9*(1), e78644. doi:10.1371/journal.pone.0078644

Zheng, D., & Gerstein, M. B. (2007). The ambiguous boundary between genes and pseudogenes: the dead rise up, or do they? *Trends in Genetics: TIG*, *23*(5), 219‑224. doi:10.1016/j.tig.2007.03.003

Zhou, X., Ren, L., Meng, Q., Li, Y., Yu, Y., & Yu, J. (2010). The next-generation sequencing technology and application. *Protein & Cell*, *1*(6), 520‑536. doi:10.1007/s13238-010-0065-3

Zmieńko, A., Samelak, A., Kozłowski, P., & Figlerowicz, M. (2014). Copy number polymorphism in plant genomes. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *127*(1), 1‑18. doi:10.1007/s00122-013-2177-7

Zou, C., Lehti-Shiu, M. D., Thibaud-Nissen, F., Prakash, T., Buell, C. R., & Shiu, S.-H. (2009). Evolutionary and Expression Signatures of Pseudogenes in Arabidopsis and Rice. *Plant Physiology*, *151*(1), 3‑15. doi:10.1104/pp.109.140632

# ANNEXES

**Liste des Annexes**

**Annexe 1** : **Additionnal data files** : **Pingault, L**., Choulet, F., Alberti, A., Glover, N., Wincker, P., The International Wheat Genome Sequencing Consortium, … Paux, E. (s. d.). Deep transcriptome sequencing provides new insights into the structural and functional organization of the wheat genome.

**Annexe 2** : International Wheat Genome Sequencing Consortium. 2014. « A Chromosome-Based Draft Sequence of the Hexaploid Bread Wheat (Triticum Aestivum) Genome ». *Science (New York, N.Y.)* 345 (6194): 1251788. doi:10.1126/science.1251788.

**Annexe 3** : Daron, J., Glover, N., **Pingault, L**., Theil, S., Jamilloux, V., Paux, E., … Choulet, F. (s. d.). Organization and Evolution of Transposable Elements along the Wheat Chromosome 3B.

**Annexe 4** : Thomas, M., **Pingault, L**., Poulet, A., Duarte, J., Throude, M., Faure, S., … Tatout, C. (s. d.). Evolutionary history of Methyltransferase 1 genes in hexaploid wheat.

**Annexe 1** : **Additionnal data files** : **Pingault, L**., Choulet, F., Alberti, A., Glover, N., Wincker, P., The International Wheat Genome Sequencing Consortium, … Paux, E. (s. d.). Deep transcriptome sequencing provides new insights into the structural and functional organization of the wheat genome.

**Table S1 :** Distribution of the 5,185 expressed protein coding genes according to their expression cluster and their chromosomal region

| File | Total reads mapped | Raw reads count | % reads mapped | Mapped | paired in sequencing | reads1 | reads2 | properly paired | % reads properly paired | with itself and mate mapped | singletons | with mate mapped to a different chr | with mate mapped to a different chr (mapQ>=5) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Roots | | | | | | | | | | | | | |
| Z10-rep1 | 1 899 394 | 90 796 198 | 2.09 | 1 899 394 | 1 899 394 | 1 026 025 | 873 369 | 1 271 186 | 66.93 | 1 310 690 | 588 704 | 31 877 | 1 793 |
| Z10-rep2 | 1 563 452 | 93 555 552 | 1.67 | 1 563 452 | 1 563 452 | 739 228 | 824 224 | 1 001 440 | 64.05 | 1 038 450 | 525 002 | 25 367 | 1 376 |
| Z13-rep1 | 1 282 168 | 186 409 424 | 0.69 | 1 282 168 | 1 282 168 | 723 447 | 558 721 | 781 463 | 60.95 | 820 896 | 461 272 | 28 198 | 1 498 |
| Z13-rep2 | 1 626 797 | 83 846 634 | 1.94 | 1 626 797 | 1 626 797 | 894 722 | 732 075 | 988 434 | 60.76 | 1 027 672 | 599 125 | 24 536 | 1 472 |
| Z39-rep1 | 2 387 994 | 84 586 562 | 2.82 | 2 387 994 | 2 387 994 | 1 241 363 | 1 146 631 | 1 319 133 | 55.24 | 1 395 764 | 992 230 | 41 770 | 2 337 |
| Z39-rep2 | 1 679 978 | 91 872 594 | 1.83 | 1 679 978 | 1 679 978 | 824 277 | 855 701 | 1 009 819 | 60.11 | 1 056 675 | 623 303 | 29 442 | 1 562 |
| Leaves | | | | | | | | | | | | | |
| Z10-rep1 | 1 388 775 | 85 280 910 | 1.63 | 1 388 775 | 1 388 775 | 745 755 | 643 020 | 884 072 | 63.66 | 913 092 | 475 683 | 21 088 | 1 788 |
| Z10-rep2 | 1 363 534 | 92 099 786 | 1.48 | 1 363 534 | 1 363 534 | 644 285 | 719 249 | 843 889 | 61.89 | 881 893 | 481 641 | 27 481 | 1 867 |
| Z23-rep1 | 1 804 799 | 102 140 692 | 1.77 | 1 804 799 | 1 804 799 | 977 439 | 827 360 | 1 161 724 | 64.37 | 1 211 607 | 593 192 | 40 477 | 2 530 |
| Z23-rep2 | 1 772 607 | 94 155 306 | 1.88 | 1 772 607 | 1 772 607 | 955 553 | 817 054 | 1 032 413 | 58.24 | 1 093 724 | 678 883 | 43 723 | 3 860 |
| Z71-rep1 | 1 816 369 | 95 116 094 | 1.91 | 1 816 369 | 1 816 369 | 979 328 | 837 041 | 1 111 429 | 61.19 | 1 175 234 | 641 135 | 55 047 | 1 908 |
| Z71-rep2 | 1 415 370 | 74 220 878 | 1.91 | 1 415 370 | 1 415 370 | 707 668 | 707 702 | 807 652 | 57.06 | 855 677 | 559 693 | 38 764 | 1 509 |
| Stems | | | | | | | | | | | | | |
| Z30-rep1 | 1 906 101 | 100 275 388 | 1.90 | 1 906 101 | 1 906 101 | 1 024 371 | 881 730 | 1 256 465 | 65.92 | 1 299 884 | 606 217 | 36 236 | 1 667 |
| Z30-rep2 | 1 863 558 | 87 877 244 | 2.12 | 1 863 558 | 1 863 558 | 998 167 | 865 391 | 1 114 637 | 59.81 | 1 173 372 | 690 186 | 36 529 | 3 876 |
| Z32-rep1 | 1 405 950 | 112 195 350 | 1.25 | 1 405 950 | 1 405 950 | 771 659 | 634 291 | 921 189 | 65.52 | 959 789 | 446 161 | 28 815 | 1 471 |
| Z32-rep2 | 1 690 052 | 82 820 236 | 2.04 | 1 690 052 | 1 690 052 | 918 077 | 771 975 | 1 038 694 | 61.46 | 1 082 086 | 607 966 | 26 645 | 2 150 |
| Z65-rep1 | 1 679 808 | 86 194 118 | 1.95 | 1 679 808 | 1 679 808 | 900 532 | 779 276 | 1 078 622 | 64.21 | 1 127 040 | 552 768 | 38 603 | 2 193 |
| Z65-rep2 | 1 513 787 | 87 322 784 | 1.73 | 1 513 787 | 1 513 787 | 749 493 | 764 294 | 893 918 | 59.05 | 948 003 | 565 784 | 38 283 | 1 872 |
| Spikes | | | | | | | | | | | | | |
| Z32-rep1 | 2 091 625 | 98 618 666 | 2.12 | 2 091 625 | 2 091 625 | 1 109 628 | 981 997 | 1 356 205 | 64.84 | 1 406 474 | 685 151 | 36 496 | 2 297 |
| Z32-rep2 | 2 115 456 | 95 541 924 | 2.21 | 2 115 456 | 2 115 456 | 1 122 017 | 993 439 | 1 331 076 | 62.92 | 1 390 057 | 725 399 | 39 143 | 2 584 |
| Z39-rep1 | 2 027 658 | 91 056 958 | 2.23 | 2 027 658 | 2 027 658 | 1 063 923 | 963 735 | 1 249 126 | 61.60 | 1 303 527 | 724 131 | 39 944 | 2 404 |
| Z39-rep2 | 2 882 872 | 104 799 124 | 2.75 | 2 882 872 | 2 882 872 | 1 505 175 | 1 377 697 | 1 724 247 | 59.81 | 1 826 055 | 1 056 817 | 69 383 | 4 481 |
| Z65-rep1 | 2 019 147 | 93 025 672 | 2.17 | 2 019 147 | 2 019 147 | 1 087 275 | 931 872 | 1 289 116 | 63.84 | 1 343 207 | 675 940 | 39 763 | 2 338 |
| Z65-rep2 | 2 294 911 | 108 141 500 | 2.12 | 2 294 911 | 2 294 911 | 1 143 404 | 1 151 507 | 1 310 792 | 57.12 | 1 392 222 | 902 689 | 48 299 | 3 238 |
| Graines | | | | | | | | | | | | | |
| Z71-rep1 | 1 873 383 | 94 392 922 | 1.98 | 1 873 383 | 1 873 383 | 1 004 710 | 868 673 | 1 235 184 | 65.93 | 1 280 778 | 592 605 | 28 705 | 2 145 |
| Z71-rep2 | 1 658 545 | 80 171 210 | 2.07 | 1 658 545 | 1 658 545 | 826 976 | 831 569 | 1 019 483 | 61.47 | 1 054 158 | 604 387 | 21 937 | 1 831 |
| Z75-rep1 | 1 389 302 | 137 089 674 | 1.01 | 1 389 302 | 1 389 302 | 716 750 | 672 552 | 872 051 | 62.77 | 916 314 | 472 988 | 35 720 | 2 649 |
| Z75-rep2 | 1 052 646 | 74 062 704 | 1.42 | 1 052 646 | 1 052 646 | 565 886 | 486 760 | 625 768 | 59.45 | 655 965 | 396 681 | 24 759 | 1 631 |
| Z85-rep1 | 1 463 618 | 99 131 804 | 1.48 | 1 463 618 | 1 463 618 | 783 886 | 679 732 | 884 078 | 60.40 | 925 514 | 538 104 | 30 318 | 1 365 |
| Z85-rep2 | 1 118 154 | 80 471 930 | 1.39 | 1 118 154 | 1 118 154 | 645 899 | 472 255 | 528 612 | 47.28 | 562 023 | 556 131 | 26 409 | 932 |

**Table S2 :** Wheat RNA samples used for RNA-seq experiments.

| Stage | Wheat growth stage | Feekes scale | Zadoks scale | Leaves | Root | Stem | Spike | Grain |
|---|---|---|---|---|---|---|---|---|
| Seedling | First leaf through coleoptile | 1 | 10 | x | x | | | |
| Three leaves | 3 leaves unfolded | | 13 | | x | | | |
| Three tillers | Main shoot and 3 tillers | | 23 | x | | | | |
| Spike at 1 cm | Pseudostem erection | 5 | 30 | | | x | | |
| Two nodes | 2nd detectable node | 7 | 32 | | | x | x | |
| Meiosis | Flag leaf ligule and collar visible | 9 | 39 | | x | | x | |
| Anthesis | 1/2 of flowering complete | | 65 | | | x | x | |
| 2 DAAs (50°C.days) | Kernel (caryopsis) watery ripe | | 71 | x | | | | x |
| 14 DAAs (350°C.days) | Medium Milk | | 75 | | | | | x |
| 30 DAAs (700°C.days) | Soft dough | | 85 | | | | | x |

**Figure S1 :** Alternative transcript length variables as functions of expression level category.

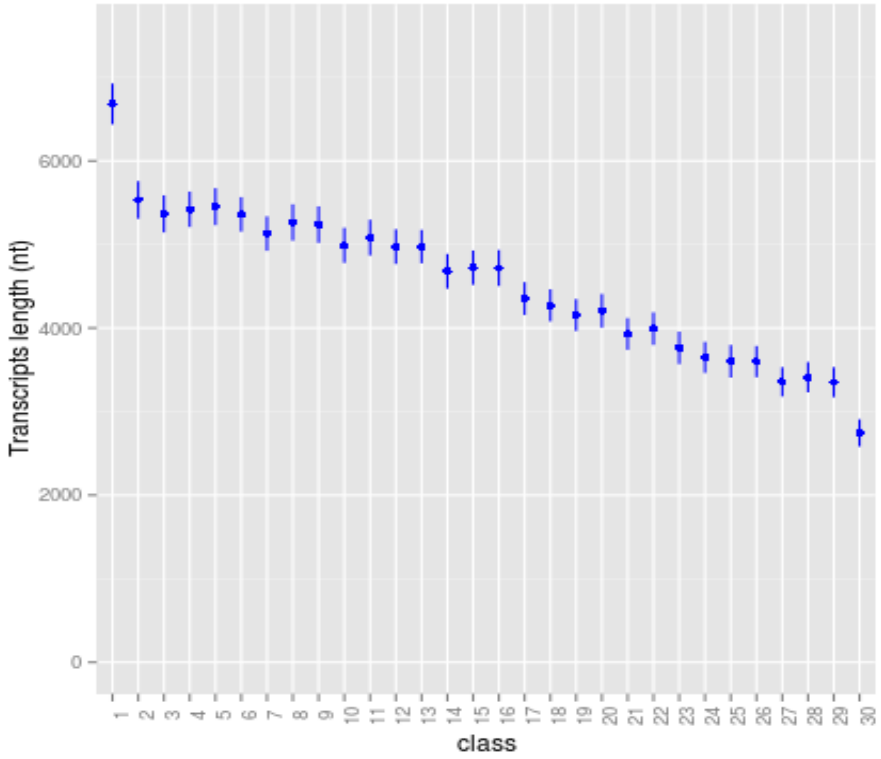


**Figure S2 :** Relationships between gene expression and gene structural and functional features in the R1 / R3 regions.

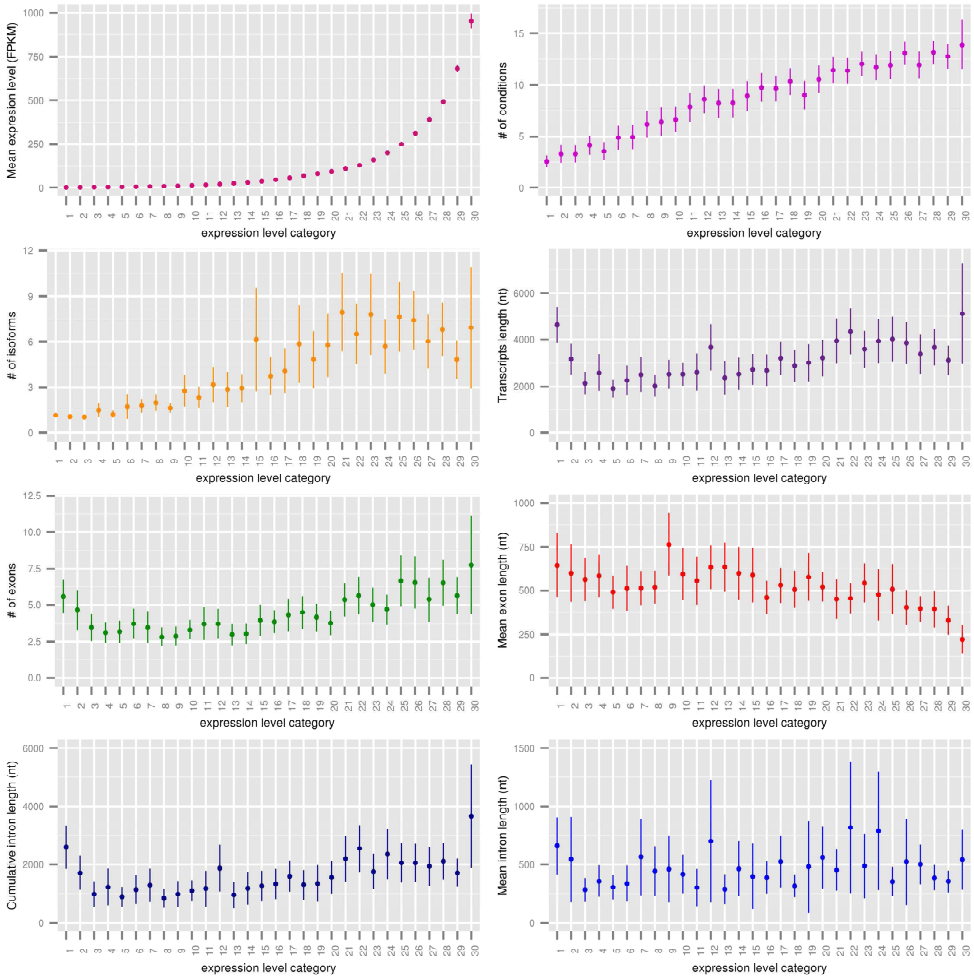
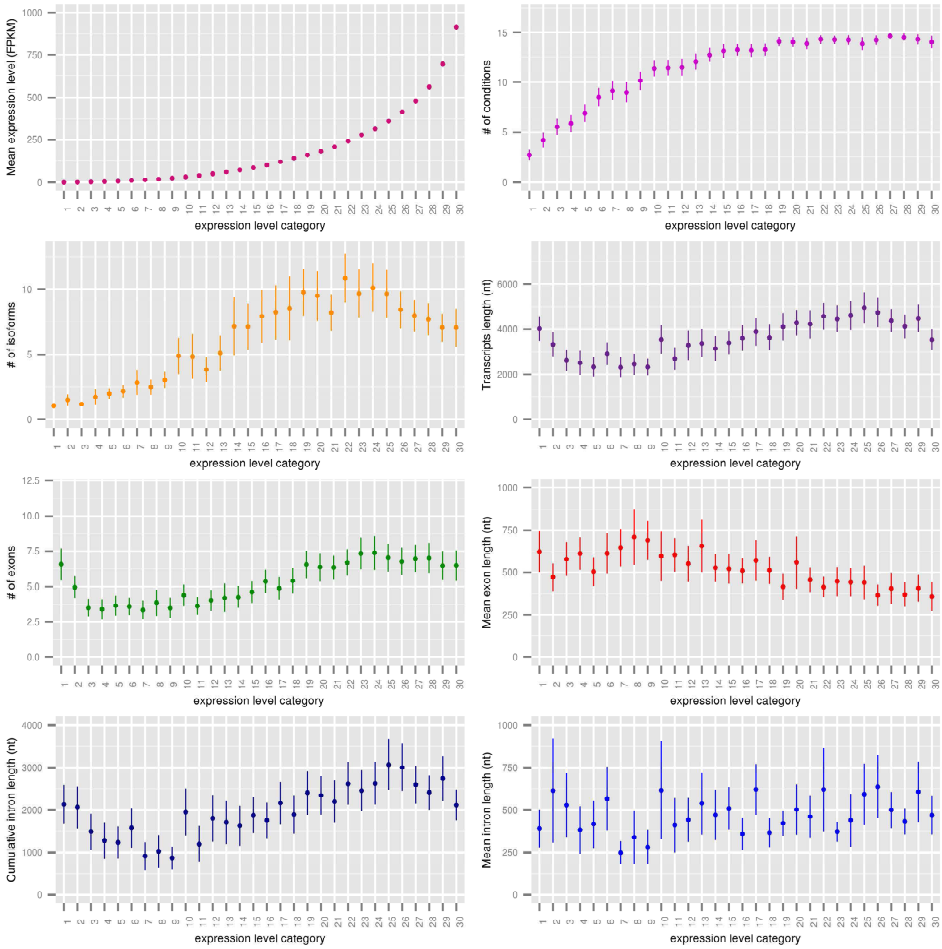


**Figure S3 : .**Relationships between gene expression and gene structural and functional features in the R2a / R2b regions.



**Annexe 2** : International Wheat Genome Sequencing Consortium. 2014. « A Chromosome-Based Draft Sequence of the Hexaploid Bread Wheat (Triticum Aestivum) Genome ». Science (New York, N.Y.) 345 (6194): 1251788. doi:10.1126/science.1251788.

WHEAT GENOME

# A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome

The International Wheat Genome Sequencing Consortium (IWGSC)*†

**An ordered draft sequence of the 17-gigabase hexaploid bread wheat (*Triticum aestivum*) genome has been produced by sequencing isolated chromosome arms. We have annotated 124,201 gene loci distributed nearly evenly across the homeologous chromosomes and subgenomes. Comparative gene analysis of wheat subgenomes and extant diploid and tetraploid wheat relatives showed that high sequence similarity and structural conservation are retained, with limited gene loss, after polyploidization. However, across the genomes there was evidence of dynamic gene gain, loss, and duplication since the divergence of the wheat lineages. A high degree of transcriptional autonomy and no global dominance was found for the subgenomes. These insights into the genome biology of a polyploid crop provide a springboard for faster gene isolation, rapid genetic marker development, and precise breeding to meet the needs of increasing food demand worldwide.**

Rich in protein, carbohydrates, and minerals, bread wheat (*Triticum aestivum* L.) is one of the world's most important cereal grain crops, serving as the staple food source for 30% of the human population. Between 2000 and 2008, wheat production fell by 5.5% primarily because of climatic trends (*1*), and, in 5 of the past 10 years, worldwide wheat production was not sufficient to meet demand (*2*). With the global population projected to exceed 9 billion by 2050, researchers, breeders and growers are facing the challenge of increasing wheat production by about 70% to meet future demands (*3*, *4*). Concurrently, growers are facing rising fertilizer and other input costs, weather extremes resulting from climate change, increasing competition between food and nonfood uses, and declining annual yield growth (*5*). A rapid paradigm shift in science-based advances in wheat genetics and breeding, comparable to the first green revolution of the 1960s, will be essential to meet these challenges. As for other major cereal crops (rice, maize, and sorghum), new knowledge and molecular tools using a reference genome sequence of wheat are needed to underpin breeding to accelerate the development of new wheat varieties.

One key factor in the success of wheat as a global food crop is its adaptability to a wide range of climatic conditions. This is attributable, in part, to its allohexaploid genome structure, which arose as a result of two polyploidization events (Fig. 1). The first of these is estimated to have occurred several hundred thousand years ago and brought together the genomes of two diploid species related to the wild species *Triticum urartu* ($2n = 2x = 14$; AA; $2n$ is the number of chromosomes in each somatic cell and $2x$ is the basic chromosome number) and a species from the Sitopsis section of *Triticum* that is believed to be related to *Aegilops speltoides* ($2n = 14$; SS) (*6*). This hybridization formed the allotetraploid *Triticum turgidum* ($2n = 4x = 28$; AABB), an ancestor of wild emmer wheat cultivated in the Middle East and *T. turgidum* sp. *durum* grown for pasta today. A second hybridization event between *T. turgidum* and a diploid grass species, *Aegilops tauschii* (DD), produced the ancestral allohexaploid *T. aestivum* ($2n = 6x = 42$, AABBDD) (*6*, *7*), which has since been cultivated as bread wheat and accounts for over 95% of the wheat grown worldwide.

With 21 pairs of chromosomes, bread wheat is structurally an allopolyploid with three homeologous sets of seven chromosomes in each of the A, B, and D subgenomes. Genetically, however, it behaves as a diploid because homeologous pairing is prevented through the action of *Ph* genes (*8*). Each of the subgenomes is large, about 5.5 Gb in size and carries, in addition to related sets of genes, a high proportion (>80%) of highly repetitive transposable elements (TEs) (*9*, *10*).

The large and repetitive nature of the genome has hindered the generation of a reference genome sequence for bread wheat. Early work focused primarily on coding sequences that represent less than 2% of the genome. Coordinated efforts generated over 1 million expressed sequence tags (ESTs), 40,000 unigenes (www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html), and 17,000 full-length complementary DNA (cDNA) sequences (*11*). These resources have enabled studies of individual genes and facilitated the development of microarrays and marker sets for targeted gene association and expression studies (*12*–*14*). At least 7000 ESTs have been assigned to chromosome-specific bins (*15*), providing an initial view of subgenome localization and chromosomal organization and facilitating low-resolution mapping of traits. More recently, high-throughput low-cost sequencing technologies have been applied to assemble the gene space of *T. urartu* (*16*) and *Ae. tauschii* (*17*), two diploid species related to bread wheat (Fig. 1). About 60,000 genic sequences were also putatively assigned to the bread wheat A, B, or D subgenomes by using assembled Illumina (Illumina, Incorporated, San Diego, CA) sequence data for *Triticum monococcum* and *Ae. tauschii* and cDNAs from *Ae. speltoides* to guide gene assemblies of five-fold whole-genome sequence reads from *T. aestivum* 'Chinese Spring' (*18*). These resources have contributed information about the genes of hexaploid wheat and its wild diploid relatives and have underpinned the

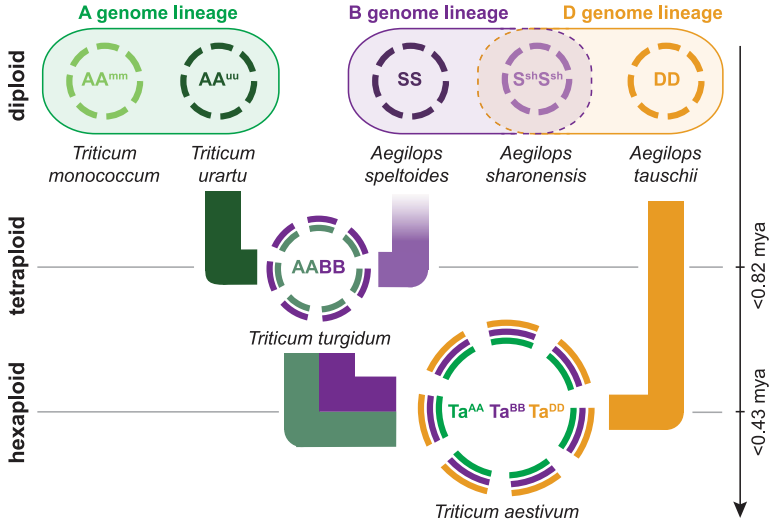



**Fig. 1. Schematic diagram of the relationships between wheat genomes with polyploidization history and genealogy.** Names and nomenclature for the genomes are indicated within circles that provide a schematic representation of the chromosomal complement for each species. Time estimates are from Marcussen *et al.* (*45*). mya, million years ago.

*All authors with their affiliations appear at the end of this paper.
†Corresponding author: K. X. Mayer (k.mayer@helmholtz-muenchen.de)

development of large sets of single-nucleotide polymorphism (SNP) markers (*19*–*21*). To date, however, relatively little is known about the position and distribution of genes on each of the bread wheat chromosomes and their evolution during the polyploidization events that resulted in the emergence of the hexaploid species.

### Survey sequencing the bread wheat genome

We used aneuploid bread wheat lines derived from double ditelosomic stocks of the hexaploid wheat cultivar Chinese Spring (*22*) to isolate, sequence, and assemble de novo each individual chromosome arm [except for 3B, which was isolated and sequenced as a complete chromosome (*23*)]. This approach reduced the complexity of assembling a highly redundant genome and enabled the differentiation of genes present in multiple copies and highly conserved homologs. Each chromosome arm, representing between 1.3 and 3.3% of the genome (*24*), was purified by flow-cytometric sorting and sequenced to a depth of between 30× and 241× with Illumina technology platforms (*25*). The paired end sequence reads were assembled with the short-read de novo assembly tool ABySS (*25*, *26*). A high proportion of reads assembled into contigs of repetitive sequence less than 200 base pairs (bp) and were excluded from the final assembly of 10.2 Gb. The quality of the assemblies and purity of chromosome arm preparations were assessed by using alignment to bin-mapped ESTs (*15*) and to the virtual barley genome (*27*). Summary statistics for the chromosome arm assemblies are shown in Tables 1 to 3. Compared with cytogenetically estimated chromosome sizes (*24*), the sequence assemblies represent 61% of the genome sequence, with the L50 of repeat-masked assemblies ranging from 1.7 to 8.9 kb.

### Repetitive DNA

We assessed the TE and sequence repeat space across the whole wheat genome and compared the repeat content of the A, B, and D subgenomes (*25*). From the frequency of mathematically defined repeats (MDRs; 20mers) (*28*), we estimated that 24 to 26% of the sequence reads contain high copy number repeats, represented by 20mers with more than 1000 copies. In total, 81% of raw reads and 76.6% of assembled sequences contained repeats, the latter showing reduced representation of Gypsy long terminal repeat (LTR) retrotransposons, as well as Mutator and Mariner-type DNA transposons.

Analysis of the distribution of transposons across the three subgenomes revealed that class I elements (retroelements) were more abundant in the A genome chromosomes relative to B or D (A > B > D), whereas class II elements (DNA transposons) showed the reverse (D > B > A). The most pronounced differences were observed between deteriorated and thus unclassifiable LTR retrotransposons, which showed a gradient of abundance across the subgenomes (A > D > B) distinct from other class I or class II elements. We hypothesize that unclassifiable LTR retrotransposons represent older (and thus more deteriorated) elements that were modified through polyploidization and ongoing TE amplification or degeneration. Assuming the amplification/degeneration dynamics are similar within each genome, the distribution of LTR retrotransposons across the three subgenomes suggest that the B genome progenitor contained a lower number of LTR retroelements and that transposon activity post-polyploidization has introduced a higher proportion of more recent amplifications into the B genome.

We observed a substantial reduction (down to 19.6%) in the TE content associated with the 0.8% (615 Mb) of the chromosomal survey sequences (CSSs) representing contigs containing high-confidence genes (for definition see below) (*25*). The analysis revealed a marked depletion of all class I elements in the neighborhood of genes, with the exception of non-LTR retrotransposons, which were enriched twofold. CACTA transposons accounted for the greatest proportion of the observed 67% reduction in class II elements, whereas minor components, especially Harbinger and miniature inverted-repeat TEs, were enriched. Selective exclusion of high-copy transposons that undergo epigenetic silencing and reduce expression by heterochromatin spreading (*29*) may result in depletion of repeat element types in the vicinity of genes.

### miRNAs

A total of 270 different putative microRNA molecules (miRNAs) (49 not previously reported)

**Table 1. Sequencing, assembly, and GenomeZipper statistics for wheat A genome chromosome arms.** Sequence indicates the total assembled sequence (>200 bp); number of contigs is after filtering of highly repetitive sequence assemblies; syntenic loci is the number of gene loci anchored to reference gene; and the last row is the number of total anchored gene loci. Blank entries in all tables indicate data not applicable; fl-cDNA, full-length cDNA; nonred., nonredundant.

| | 1AS | 1AL | 2AS | 2AL | 3AS | 3AL | 4AS | 4AL | 5AS | 5AL | 6AS | 6AL | 7AS | 7AL | Σ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Assembly* | | | | | | | | | | | | | | | |
| Chromosome size (Mbp) | 275 | 523 | 391 | 508 | 360 | 468 | 317 | 539 | 295 | 532 | 336 | 369 | 407 | 407 | 5,727 |
| Sequence (Mbp) | 178.1 | 250 | 255.2 | 328.2 | 201.8 | 247.2 | 282.3 | 362 | 198.8 | 318.1 | 219.2 | 214.4 | 198 | 252.4 | 3,505.7 |
| Coverage (*x*-fold) | 0.65 | 0.48 | 0.65 | 0.65 | 0.56 | 0.53 | 0.89 | 0.67 | 0.67 | 0.60 | 0.65 | 0.58 | 0.49 | 0.62 | 0.62 |
| L50 (bp) | 2,242 | 2,639 | 2,398 | 2,688 | 1,404 | 1,346 | 2,782 | 3,053 | 3,509 | 2,078 | 2,669 | 2,154 | 1,470 | 2,271 | |
| *Repeat* | | | | | | | | | | | | | | | |
| No. of contigs | 34,793 | 26,746 | 34,722 | 45,893 | 33,943 | 43,823 | 32,079 | 64,364 | 19,719 | 47,572 | 28,041 | 34,030 | 44,175 | 35,586 | 542,486 |
| L50 | 4,769 | 6,369 | 6,678 | 6,677 | 3,846 | 3,789 | 7,499 | 6,601 | 8,713 | 5,355 | 7,091 | 6,589 | 4,397 | 5,849 | |
| *GenomeZipper* | | | | | | | | | | | | | | | |
| No. of markers | 147 | 380 | 139 | 278 | 106 | 332 | 167 | 200 | 150 | 309 | 174 | 286 | 169 | 278 | 3,115 |
| No. of wheat fl-cDNAs | 95 | 241 | 162 | 258 | 134 | 240 | 153 | 189 | 54 | 231 | 94 | 181 | 178 | 155 | 2,365 |
| No. of nonred. contigs | 937 | 1,750 | 1,673 | 2,499 | 1,323 | 2,300 | 848 | 2,613 | 574 | 2,495 | 811 | 1,422 | 2,100 | 1,600 | 22,945 |
| No. of syntenic gene loci | 544 | 1,515 | 1,155 | 1,816 | 850 | 1,628 | 842 | 1,642 | 405 | 1,821 | 647 | 1,073 | 1,228 | 1,049 | 16,215 |
| No. of anchored gene loci | 649 | 1,811 | 1,262 | 2,032 | 929 | 1,864 | 948 | 1,777 | 522 | 2,050 | 794 | 1,279 | 1,349 | 1,269 | 18,535 |
| *POP-Seq Positioning* | | | | | | | | | | | | | | | |
| No. of contigs | 38,940 | 45,649 | 34,853 | 32,941 | 31,094 | 49,586 | 25,068 | 27,248 | 5,578 | 35,333 | 28,234 | 30,828 | 31,628 | 32,435 | 449,415 |
| No. of anchored gene loci | 972 | 1,720 | 1,452 | 1,913 | 788 | 1,302 | 883 | 1,702 | 137 | 1,579 | 1,145 | 1,305 | 1,305 | 1,094 | 17,297 |
| No. of anchored gene loci | 618 | 1,257 | 1,408 | 1,903 | 769 | 1,469 | 778 | 1,116 | 678 | 2,432 | 995 | 1,458 | 1,405 | 1,711 | 17,997 |

were identified corresponding to 98,068 predicted miRNA-coding loci (25). Only 1668 loci (1.7%) evidenced expression on the basis of publicly available ESTs and of RNA sequencing (RNA-seq) data reported in this work, consistent with previous analyses in wheat (30, 31). Similarly, we observed that class II DNA transposons, specifically TcMar transposons, were predominantly found in miRNAs. For 87% of the putative miRNA-coding loci, at least one putative target gene was identified in the wheat CSS. A total of 6615 predicted miRNA-coding sequences (44 with evidence of expression) were characterized by at least one mature sequence and one target site covered by the same repeat element. This suggests that an active miRNA could arise when an advantageous regulatory niche evolves from a series of random

**Table 2. Sequencing, assembly, and GenomeZipper statistics for wheat B genome chromosome arms.** Sequence indicates the total assembled sequence (>200 bp); number of contigs is after filtering of highly repetitive sequence assemblies; syntenic loci is the number of gene loci anchored to reference gene; and the last row is the number of total anchored gene loci.

| | 1BS | 1BL | 2BS | 2BL | 3B | 4BS | 4BL | 5BS | 5BL | 6BS | 6BL | 7BS | 7BL | Σ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | *Assembly* | | | | | | | |
| Chromosome size (Mbp) | 314 | 535 | 422 | 506 | 993 | 391 | 430 | 290 | 580 | 415 | 498 | 360 | 540 | 6,274 |
| Sequence (Mbp) | 212.8 | 299.4 | 292 | 404.5 | 638.6 | 308.2 | 248.7 | 174.5 | 415.2 | 210.2 | 257.4 | 206.1 | 259.6 | 3,927.2 |
| Coverage (*x*-fold) | 0.68 | 0.56 | 0.69 | 0.80 | 0.64 | 0.79 | 0.58 | 0.60 | 0.72 | 0.51 | 0.52 | 0.57 | 0.48 | 0.63 |
| L50 (bp) | 3,287 | 3,120 | 3,711 | 2,941 | 2,655 | 3,463 | 1,974 | 3,315 | 2,924 | 2,366 | 2,031 | 2,428 | 1,556 | |
| | | | | | | | *Repeat* | | | | | | | |
| No. of contigs | 26,050 | 29,783 | 35,743 | 75,879 | 75,022 | 38,515 | 46,576 | 18,001 | 75,887 | 29,566 | 35,727 | 24,119 | 58,554 | 569,422 |
| L50 | 7,413 | 7,151 | 8,069 | 6,890 | 6,855 | 8,755 | 5,883 | 7,365 | 7,537 | 4,972 | 4,824 | 6,435 | 4,144 | |
| | | | | | | | *GenomeZipper* | | | | | | | |
| No. of markers | 78 | 348 | 278 | 428 | 500 | 46 | 145 | 167 | 404 | 217 | 245 | 140 | 198 | 3,194 |
| No. of wheat fl-cDNAs | 78 | 219 | 155 | 268 | 479 | 97 | 170 | 66 | 360 | 88 | 147 | 109 | 137 | 2,373 |
| No. of nonred. contigs | 776 | 1,927 | 1,859 | 3,079 | 5,011 | 893 | 1,634 | 576 | 3,296 | 915 | 1,525 | 1,172 | 1,890 | 24,553 |
| No. of syntenic gene loci | 485 | 1,485 | 1,181 | 1,973 | 3,123 | 788 | 1,155 | 426 | 2,315 | 565 | 1,003 | 733 | 1,050 | 16,282 |
| No. of anchored gene loci | 546 | 1,745 | 1,388 | 2,265 | 3,490 | 819 | 1,243 | 565 | 2,600 | 728 | 1,177 | 838 | 1,203 | 18,607 |
| | | | | | | | *POP-Seq Anchoring* | | | | | | | |
| No. of contigs | 31,038 | 50,219 | 33,603 | 54,522 | 99,341 | 50,927 | 41,135 | 19,794 | 49,140 | 30,962 | 38,064 | 48,514 | 50,397 | 597,656 |
| No. of anchored gene loci | 956 | 1,881 | 1,588 | 2,389 | 3,772 | 1,365 | 1,433 | 727 | 2,857 | 831 | 996 | 1,055 | 1251 | 21,101 |

**Table 3. Sequencing, assembly, and GenomeZipper statistics for wheat D genome chromosome arms.** Sequence indicates the total assembled sequence (>200 bp); number of contigs is after filtering of highly repetitive sequence assemblies; syntenic loci is the number of gene loci anchored to reference gene; and the last row is the number of total anchored gene loci.

| | 1DS | 1DL | 2DS | 2DL | 3DS | 3DL | 4DS | 4DL | 5DS | 5DL | 6DS | 6DL | 7DS | 7DL | Σ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | *Assembly* | | | | | | | |
| Chromosome size (Mbp) | 224 | 381 | 317 | 412 | 321 | 450 | 232 | 417 | 259 | 491 | 324 | 389 | 381 | 347 | 4,937 |
| Sequence (Mbp) | 128.2 | 254.4 | 166 | 261.6 | 145.4 | 186.5 | 142.1 | 347.6 | 148 | 236.8 | 156.6 | 199.8 | 209.1 | 222.9 | 2,805 |
| Coverage (*x*-fold) | 0.57 | 0.67 | 0.52 | 0.63 | 0.45 | 0.41 | 0.61 | 0.83 | 0.57 | 0.48 | 0.48 | 0.51 | 0.55 | 0.64 | 0.57 |
| L50 (bp) | 2,850 | 2561 | 1241 | 701 | 515 | 967 | 3278 | 1013 | 2,353 | 2,647 | 4,297 | 2,077 | 1,967 | 3,638 | |
| | | | | | | | | *Repeat* | | | | | | | |
| No. of contigs | 17,725 | 35,770 | 43,044 | 110,446 | 46,795 | 69,259 | 18,245 | 197,398 | 22,449 | 34,622 | 16,077 | 26,236 | 36,701 | 26,737 | 701,504 |
| L50 | 6,622 | 6,297 | 4,635 | 3,247 | 1,697 | 2941 | 7428 | 1855 | 5945 | 7049 | 8,904 | 6,821 | 5,031 | 7,399 | |
| | | | | | | | | *GenomeZipper* | | | | | | | |
| No. of markers | 258 | 653 | 457 | 739 | 379 | 633 | 269 | 498 | 225 | 744 | 297 | 411 | 579 | 515 | 6,657 |
| No. of wheat fl-cDNAs | 89 | 251 | 177 | 323 | 128 | 244 | 130 | 255 | 99 | 375 | 103 | 208 | 200 | 212 | 2,794 |
| No. of nonred. contigs | 968 | 2,797 | 3,023 | 5,804 | 2,933 | 3,712 | 1,231 | 3,174 | 890 | 3,436 | 973 | 1,923 | 3,006 | 2,083 | 35,953 |
| No. of syntenic gene loci | 474 | 1,483 | 1,197 | 2,141 | 799 | 1,575 | 779 | 1,277 | 454 | 2,073 | 538 | 1,117 | 1,222 | 1,099 | 16,228 |
| No. of anchored gene loci | 642 | 1,882 | 1,475 | 2,542 | 1,051 | 1,923 | 912 | 1,582 | 598 | 2,482 | 758 | 1,347 | 1,592 | 1,423 | 20,209 |
| | | | | | | | | *POP-Seq Anchoring* | | | | | | | |
| No. of contigs | 7,686 | 24,149 | 24,652 | 31,359 | 26,447 | 37,874 | 14,198 | 23,842 | 14,458 | 29,604 | 18,701 | 23,763 | 41,796 | 31,832 | 350,361 |

TE insertions and may represent a means by which a network of putative miRNAs and target genes may develop, even before miRNA activation (*32*).

### Protein-coding genes

Annotation of protein-coding gene sequences in the CSS assemblies had its basis in comparisons to annotated genes in related grasses [*Brachypodium distachyon* (*33*), *Oryza sativa* (*34*), *Sorghum bicolor* (*35*), and *Hordeum vulgare* (*27*)], as well as publically available wheat full-length cDNAs (fl-cDNAs) (*11*) and RNA-seq data generated from five tissues of a Chinese Spring cultivar at three different developmental stages. Briefly, the reference grass coding sequences and wheat transcript resources were mapped separately to assembled CSS contigs, and the alignments were merged to define the exact coordinates of gene loci, alternative splicing forms, and transcripts with no similarity to related grass genes (*25*).

This analysis identified 976,962 loci with 1,265,548 distinct splicing variants. A total of 133,090 loci showing homology to related grass genes were classified as high confidence (HC) gene calls. These were further subdivided into four groups (HC1 to HC4) on the basis of the proportion of the length of the reference gene covered by a predicted locus. Of these, 124,201 (93.3%) genes were annotated on individual chromosome arm sequences, and the remaining 6.7% corresponded to wheat transcripts, which were not detected in the CSS assemblies (Fig. 2A). In total, 55,249 (44%) of the loci assigned to chromosomes were classified as HC1, that is, representing functional genes spanning at least 70% of the length of the supporting evidence (Table 4). The remaining 56% of HC genes comprised genes that were fragmented in the assembly and thus could only be partially structurally defined or were classified as gene fragments and pseudogenes. We expect that many of these will be merged as further sequencing improves the coverage and quality of genic sequences. On the basis of the level of completion of the assembly and the detection rate of HC1 genes (*25*), we estimated that the wheat genome contains 106,000 functional protein-coding genes. This supports gene number estimates ranging between 32,000 and 38,000 for each diploid subgenome in hexaploid wheat and is consistent with findings in related diploid species (*16*–*18*, *20*, *36*).

Consistent with observations of high levels of non–protein-coding loci in both plants (*27*, *37*) and animals (*38*), 890,576 loci did not share any, or only low, similarity with related grass genes. Loci with low sequence similarity (88,998) were defined as low-confidence (LC) genes, and the remainder were classified as repeat-associated, noncoding, or non–homology-supported loci (*25*). More than 96% of public wheat ESTs (HarvEST) mapped to the CSS gene sets (BLASTN; *E* value $<10^{-10}$), including 89% that correspond to HC gene-coding loci, demonstrating that the CSS assemblies contain a high representation of the current gene inventory of the bread wheat genome.

Our analysis revealed that 49% of the HC genes exhibit alternative splicing (AS) with an average of 2.6 transcripts per locus. This may be an underestimation, because 69% of the most complete gene loci (HC1) were alternatively spliced with an average of 3.5 transcripts per locus. Evidence that additional AS variants will be identified has already emerged from a preliminary assessment of gene structure prediction using proteomics analyses. In a study of 63 genes, 50 (81%) structures were confirmed, 8 (13%) provided evidence for alternative gene structures, whereas 5 were absent in the structural gene calls. Extrapolating these data to the whole genome, we estimate that hexaploid bread wheat encodes more than 300,000 distinctive protein-coding transcripts. The proportion of genes exhibiting AS appeared to be similar in all three subgenomes and is consistent with the transcriptional complexity reported for plant species such as *Arabidopsis thaliana* (*39*) and *H. vulgare* (*27*).

### Gene distribution and order

Analysis of the gene distribution across the three subgenomes revealed a higher number of gene loci on the B subgenome (44,523; 35%) compared with the A and D subgenomes, which contained 40,253 (33%) and 39,425 (32%), respectively (Fig. 2A). This distribution was not consistent at the chromosomal level. For example, the gene distribution across homeologous group 3 chromosomes is 30% 3A, 42% 3B, and 28% 3D, whereas in homeologous group 7 the D genome contains the highest proportion of genes. These observations may reflect preexisting differences in the subgenomes before polyploidization or indicate that drivers determining the composition of the genome do not act at the subgenome level but regionally.

Up to 2.4-fold variation in gene density was observed on different chromosome arms, ranging from 4.4 loci per Mb (5AS) up to 10.4 loci per Mb (2DL) (Fig. 2B). Consistent with observations in rye (*40*) and the complete sequence of wheat chromosome 3B (*23*), on average 53.2% of the

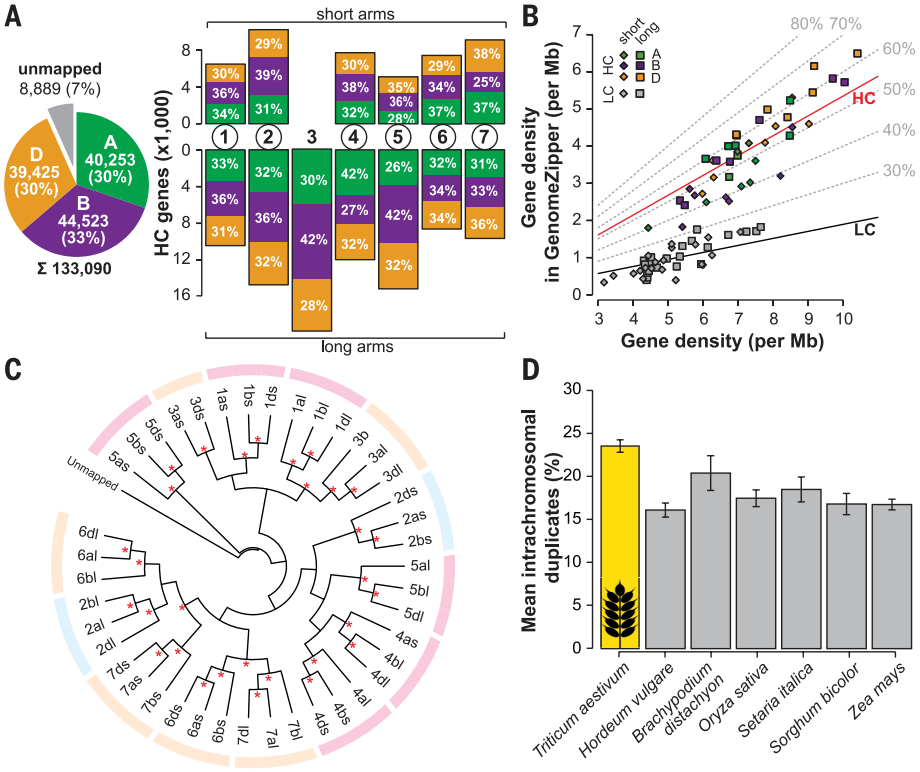



**Fig. 2. Gene content, density, synteny, structural conservation, and tandemly duplicated genes.** (**A**) Total number of HC bread wheat genes identified on the A (green), B (purple), and D (orange) subgenomes (left) and their distribution on individual chromosome arms or chromosomes (in the case of group 3) (right). (**B**) Syntenic conservation of HC and LC genes for each chromosome arm defined by the ratio of the number of genes anchored in the GenomeZipper and the number of annotated genes normalized per Mb of physical chromosome(-arm) size. Solid lines visualize average syntenic conservation for LC (black) and HC (red) genes, and dashed lines give isochores for different percentages of synteny. (**C**) Conservation of gene family composition between single chromosome arms. Color-coding in the outer ring indicates relatedness of the respective branches (A/D > B, light orange; A/B > D, light blue; B/D > A, light red). Red asterisks mark edges with boot-strapping values > 0.95. (**D**) Proportion of lineage-specific, intrachromosomally duplicated genes in the wheat genome compared with other grass genomes. Error bars indicate deviations among individual chromosomes.

**Table 4. Characteristics of HC bread wheat genes.** Distinct exons means that exons of two or more transcripts were counted once if they had identical start and stop positions; mean transcripts and mean exons are transcripts per locus and exons per locus, respectively; the second mean exons row shows exons per transcript.

| | HC1 | HC2 | HC3 | HC4 | Σ |
|---|---|---|---|---|---|
| Gene loci | 55,249 | 14,367 | 15,475 | 39,110 | 124,201 |
| Single exon | 9,181 (17%) | 3,230 (22%) | 4,906 (32%) | 20,375 (52%) | 37,692 (30%) |
| Multiple exon | 46,068 (83%) | 11,137 (78%) | 10,569 (68%) | 18,735 (48%) | 86,509 (70%) |
| Alternatively spliced | 38,059 (69%) | 7,916 (55%) | 6,465 (42%) | 8,728 (22%) | 61,168 (49%) |
| Mean size (bp) | 3,319 | 2,204 | 1,608 | 901 | 2,216 |
| Transcripts | 194,624 | 37,116 | 31,957 | 61,450 | 325,147 |
| Mean transcripts | 3.52 | 2.58 | 2.07 | 1.57 | 2.62 |
| Distinct exons | 538,250 | 94,864 | 74,630 | 117,530 | 825,274 |
| Mean exons | 9.74 | 6.60 | 4.82 | 3.01 | 6.64 |
| Mean exons[3] | 6.29 | 4.45 | 3.52 | 2.56 | 5.1 |
| Mean size (bp) | 321 | 315 | 314 | 281 | 314 |

HC genes were located on syntenic chromosomes compared to *B. distachyon* (Bd), *O. sativa* (Os), and *S. bicolor* (Sb). The average level of synteny for genes located on the D genome chromosomes (58%) was higher than the average for those on the A (51%) and the B (50%) chromosomes. Sequence conservation in LC genes is low, and, in comparison to HC genes, reduced syntenic conservation is observed. Thus, although the majority of LC genes are likely to result from the frequent generation of gene fragments by double-strand repair mechanisms or are deteriorated (pseudo)genes that were fragmented after the divergence from the other sequenced grass genomes (*10*), the retained synteny to other grass genomes suggests that some LC genes may be functional.

To determine the extent of gene conservation across homeologous chromosomes, we clustered the HC genes into protein families by sequence similarity (Fig. 2C) (*25*). With the exception of chromosome 4AL, the genes on all chromosome arms clustered with their corresponding homologs. The pattern of clustering observed for 4A is consistent with a known pericentromeric inversion and two translocations of segments from chromosome arms 5AL and 7BS (*41*, *42*). All possible cluster topologies were found between genes on the A, B, and D genomes. Overall, the patterns of conservation suggest that the gene content of the A and B homeologous chromosomes is more similar to the D genome chromosomes than to each other. This observation contradicts a model of bifurcating evolutionary relationships between the A, B, and D genomes but is consistent with models of interlineage hybridization (i.e., reticulate evolution) in the Triticeae (*43*, *44*) and corroborate phylogenomic analyses that suggest that the D genome is a product of homoploid hybrid speciation between A and B genome ancestors >5 million years ago (*45*). Although the potential for preexisting differences needs to be considered, the preservation of gene copies in each of the A, B, and D genomes provides evidence for their structural autonomy, a likely consequence of independent pairing during meiosis (*46*). A high degree of subgenome autonomy was also reflected in the observed patterns of gene expression (see below).

We used two independent but complementary approaches to generate an order for the many small contigs that comprise the chromosome arm assemblies (*25*). The GenomeZipper approach (*47*) combines the syntenic conservation of gene order in grasses (*48*) and the known gene orders of fully sequenced grass genomes (*33*–*35*) with high-density SNP-based genetic maps (*21*, *49*) to create a virtual gene order in wheat. The number of genes anchored per chromosome (chr.) ranged from 2125 (chr. 6B) to 4404 (chr. 2D) (Table 1). Overall, the GenomeZipper inferred positions of 21,221, 22,051, and 22,813 genes, respectively, in the A, B, and D genomes. To complement this, the POPSEQ approach (*50*) was used to build an ultradense genetic map comprising 13.3 million SNPs identified after shallow-coverage whole-genome sequencing of 90 doubled haploid individuals of the synthetic W7984 × Opata M85 population (*51*). This map assigned a partially overlapping set of 17,297, 21,101, and 17,997 HC genes, respectively, to the individual chromosomes of the A, B, and D genomes. The POPSEQ genetic map showed concordance with the gene assignments to flow-sorted chromosomes (99.4%) and the GenomeZipper (99.8%). The two inferred gene orders along chromosomes were also largely collinear (Spearman's correlation coefficient = 0.85). From both anchored data sets, we were able to position a nonredundant set of 75,183 HC genes on the 21 chromosomes of bread wheat by genetic mapping and/or syntenic conservation.

Gene duplication is frequently observed in plant genomes, arising from polyploidization or through tandem or segmental duplication associated with replication (*52*). For each wheat chromosome, the percentage of genes that have undergone lineage-specific intrachromosomal duplication was determined with OrthoMCL (*53*). By using the HC1 genes, we estimated that between 19.1% (chr. 7B) and 29.7% (chr. 2B) (23.6% average for all chromosomes) of the genes are duplicated on each chromosome (*25*). Comparison of the number of duplicated genes identified by this analysis for chr. 3B (25.3% of HC1 genes) with the 3B reference pseudomolecule (37% duplicated genes) (*23*) indicated that we are likely underestimating the number of duplicated genes. This is due to the fragmented nature of the assemblies obtained from whole-genome or chromosome-shotgun sequences that collapse highly conserved duplicates. No significant differences in the proportion of duplicates were observed between the three subgenomes ($\chi^2$ test, $\chi^2 = 3.8$, $P = 0.15$).

For each chromosome, an average of 73% of the duplicates are located on one of the chromosome arms, suggesting that they may be tandem duplicates that arise through unequal crossing-over and replication-dependent chromosome breakage (*54*) or through the activity of transposable elements. When compared with the percentage of intrachromosomal duplicates found in rice, sorghum, barley, maize, and foxtail millet (17 to 20%) (*27*, *33*–*35*, *55*, *56*), the proportion of gene duplications in wheat was significantly higher (Fig. 2D; Tukey's honest significant difference, pairwise $P < 0.007$).

## Comparisons with related species

We assembled sequence data from seven species related to progenitors of the bread wheat A, B, and D subgenomes (*25*). Illumina whole-genome sequence data and assemblies were generated from two tetraploid wheat cultivars (AABB) *T. turgidum* 'Cappelli' (originating from Italy) and *T. turgidum* 'Strongfield' (originating from Canada) as well as from the diploid genome of *Ae. speltoides* (SS). These data were combined with whole-genome sequence data from *T. urartu* ($AA^{uu}$) (*16*), *T. monococcum* ($AA^{mm}$), *Ae. tauschii* (DD) (*17*), and *Aegilops sharonensis* ($S^{sh}S^{sh}$). For the unannotated genomes of *T. turgidum*, *T. monococcum*, *Ae. speltoides*, and *Ae. sharonensis*, proteins of annotated grass genomes (*27*, *33*, *35*, *57*) and *T. aestivum* gene models were projected on the sequence assemblies.

Genes and gene families in the hexaploid, tetraploid, and diploid genomes were then compared to assess the dynamics of gene retention or loss after polyploidization and to define the core wheat genes. When comparing the sizes of gene families in *Ae. tauschii* (*17*) and *T. urartu* (*16*) diploid genomes with the individual subgenomes of hexaploid wheat (Fig. 3, A and B), we found that gene loss mainly affected genes belonging to expanded families, consistent with previous observations (*18*). In contrast, singletons (i.e., genes without paralogous copies within the same genome) were not usually subject to gene loss after polyploidization. Pronounced variations of gene copy retention or loss patterns were observed depending on the gene family considered.

Highly similar gene retention rates were found for all bread wheat subgenomes in comparison to *Ae. tauschii* and *T. urartu* [0.91 (A), 0.94 (B), and 0.89 (D) versus *Ae. tauschii* and 0.91 (A), 0.96 (B), and 0.91 (D) versus *T. urartu*] (Fig. 3, A and B). The extent of gene loss in the D subgenome, the most recent addition to the hexaploid genome, appeared slightly lower than the more ancient A and B subgenomes. Thus, as observed for

the gene content and structural similarities between individual chromosome arms, we found no evidence for a gradual gene loss induced by polyploidization. This may indicate that gene loss occurred rapidly after polyploid formation, followed by stabilization of gene content consistent with observations in newly created polyploids (*58*, *59*) and gene retention in cotton (*60*).

We conducted a clustering analysis of gene families and determined the number of genes in the bread wheat subgenomes that have an ortholog in the genomes from the A genome lineage (*T. urartu* and *T. monococcum*), the closest known relatives for the B lineage (*Ae. sharonensis* and *Ae. speltoides*), the D lineage (*Ae. tauschii*), as well as in the tetraploid *T. turgidum* genome (Fig. 3C). We found that the A, B, and D subgenomes contain very similar proportions of genes (60.1 to 61.3%) with orthologs in all the related diploid genomes. We also estimated the contribution of unique genes of the three subgenomes to the bread wheat genome. Because the absence of a particular gene in a single species could be due to incomplete sequence coverage or assembly errors, only lineage-specific gene family absence was considered in the analysis. Only a small fraction of the genes (1.3 to 1.7%) were specific to the A, B, or D lineages, demarcating the likely upper estimate of unique genes or gene families added to the bread wheat gene complement by the individual subgenomes.

High sequence similarity between genes in the bread wheat subgenomes impedes efficient marker development and the identification of nonsynonymous sequence variations that can potentially affect gene or protein functionality. We delineated single-nucleotide variations (SNVs) between the bread wheat genes and the diploid and tetraploid related genomes and reconstructed phylogenetic relationships by using unrooted parsimony (Fig. 4A) (*25*). In total, 11,435 SNVs within 6498 genes were specific to bread wheat and thus have likely been introduced after the second polyploidization event. Although most relationships support the known phylogeny of wheat, *Ae. sharonensis* was placed closer to the bread wheat D subgenome and *Ae. tauschii* than to *Ae. speltoides* and the B genome branch. This suggests that the Sitopsis group, which includes *Ae. sharonensis* and *Ae. speltoides*, is deeply furcated and related to both D and B genome branches.

The potential impact of all SNVs detected on proteins was measured by using Grantham amino acid substitution matrix scores (*25*, *61*). Most of the substitutions (80.8%) in gene sequences were conservative or moderately conservative and were randomly distributed across all chromosomes. However, bread wheat genes contained a higher proportion of substitutions with a predicted large impact on the protein functionality (i.e., moderately radical and radical changes) compared with their closest diploid or tetraploid relatives. This points to gene redundancy in hexaploid bread wheat enabling accelerated sequence evolution and potentially the evolution of novel protein functions.

We used the bread wheat gene annotation to analyze the introduction of likely premature stop codons in diploid and tetraploid related genomes as a measure for the rate and degree of pseudogenization (Fig. 4B). Using only the highest confidence genes (HC1), 290 (1.6%; *T. turgidum* A genome versus *T. aestivum* A genome) to 636 (3.6%; *Ae. sharonensis* versus *T. aestivum* D genome) gene loci had characteristics of pseudogenization in the respective related diploid genomes compared with the respective bread wheat A, B, and D subgenomes. Most of these likely pseudogenized loci were specific to the respective genomes, although overlapping candidate pseudogenized loci were also observed. However, the numbers of genes in these categories were small, ranging from 0.1 to 0.7%. Similar inferred pseudogenization rates were found in the A and B subgenomes of *T. turgidum* [290 (1.6%) in the A genome and 395 (2.0%) in the B genome, respectively], indicating no preferential pseudogenization or gene loss in any of the subgenomes. The number of pseudogenes observed in the D genome was similar to that of the A and B subgenomes and their diploid relatives, suggesting a rapid elimination process for pseudogenes. These findings are consistent with those from other plants, notably among *Arabidopsis* ecotypes (*62*), and smaller-scale analysis of pseudogenization dynamics within the bread wheat genome (*63*).

Earlier studies showed a high degree of gene sequence similarity between A, B, and D bread wheat subgenomes and their related diploid species (*6*). We analyzed the sequence conservation in bread wheat chromosomes compared to their diploid and tetraploid relatives to test for intergenomic translocations or introgressions (Fig. 4C). The sequences of genes were highly conserved, exceeding 99% identity, between the hexaploid subgenomes and their respective diploid relatives. High levels of conservation, averaging 97%, were also found between the A, B, and D lineages.

No gradients in sequence conservation were apparent along the chromosomes for the most closely related genomes. However, when comparing more distant genomes (e.g., *T. aestivum* D genome versus *T. urartu*), higher levels of sequence conservation were observed in genes located in proximal, pericentromeric, and centromeric regions. These results are consistent with findings for the 3B pseudomolecule analysis that demonstrated a partitioning of the chromosome with variable telomeric regions and a more conserved central chromosomal region (*23*). The most pronounced deviation in gene sequence similarity from the overall distribution is found for chr. 4A, which has undergone a recent inversion and translocations from chrs. 5A and 7B (*41*, *42*) (Fig. 4C). Other, smaller regions showing altered similarity profiles were also observed on other chromosomes (e.g., chrs. 2A and 7B) (*25*) suggesting the presence of further small translocations or introgressions that may have occurred after hybridization.

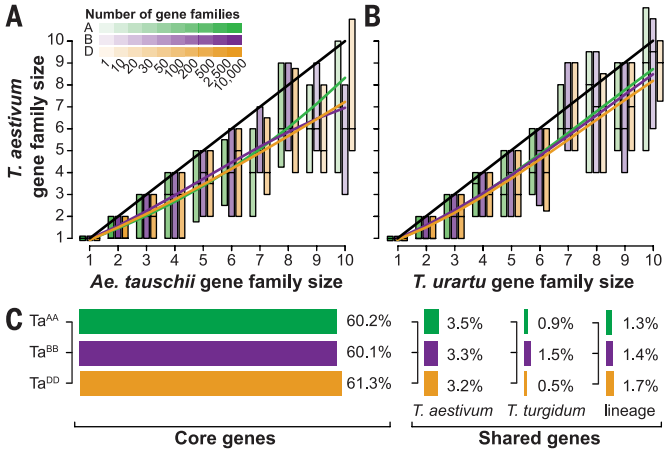



**Fig. 3. Gene conservation and the wheat pan- and core genes.** (**A** and **B**) Relationship between gene family sizes in diploid *Ae. tauschii* (A) and *T. urartu* (B) and each subgenome of hexaploid bread wheat (colors as in Fig. 2A). Boxes visualize the lower and upper quartiles of gene family sizes. Color intensity indicates the number of gene families in the respective bin. The black line shows a 1:1 gene copy number relationship for bread wheat, *Ae. tauschii*, and *T. urartu*, and colored lines show the regression fit for observed gene family size in the wheat subgenomes. (**C**) Percentages of genes of the bread wheat subgenomes that show significant sequence similarity to other genomes: Core genes correspond to genes with hits to all subgenomes as well as to *T. turgidum* and all diploid related progenitor genomes; shared genes–*T. aestivum* are genes with hits to any other *T. aestivum* subgenome but not to *T. turgidum* or any of the closest diploid relatives; shared genes–*T. turgidum* correspond to genes with hits to *T. turgidum* but not to any of the closest diploid relatives; shared genes–lineage, with hits to the subgenome's closest relative genome but not to *T. turgidum* or any of the other closest related genomes.

## Hexaploid genome phylogeny

To further test the relatedness of the A, B, and D subgenomes across the entire wheat genome, we used syntenic gene alignments to estimate maximum likelihood phylogenetic trees. We obtained 2269 trees and analyzed them for topological variation. Across all chromosome groups, 40, 35, and 25% of the gene phylogenies supported AD,

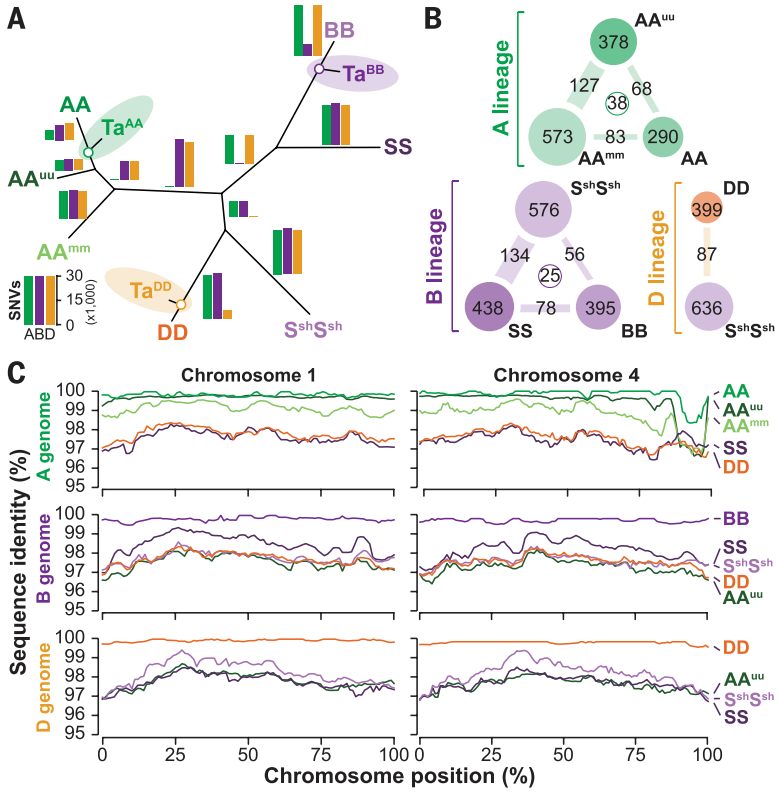



**Fig. 4. Molecular evolution of the wheat lineage.** SNVs were identified for coding sequences of bread wheat genes (TaAA, TaBB, and TaDD) against diploid *T. monococcum* ($AA^{mm}$), *T. urartu* ($AA^{uu}$), *Ae. speltoides* (SS), *Ae. sharonensis* (SshSsh), *Ae. tauschii* (DD), and tetraploid *T. turgidum* (AABB). (**A**) Unrooted phylogeny constructed on the basis of SNVs between bread wheat and its diploid or tetraploid relatives. The respective number of SNVs in each phylogenetic internodes is indicated with bar charts (scale at bottom left corner); colors indicate the respective bread wheat subgenome as in Fig. 2A. (**B**) Genes with stop codons in the respective related diploid genomes in comparison to the bread wheat A, B, and D subgenomes. Numbers in node connectors or in the center correspond to the number of introduced stop codons found in two (node connectors) or all (center) related genomes. (**C**) Chromosomal distribution of sequence identity between bread wheat genes and the diploid and tetraploid relatives for homeologous chromosomes.

BD, and AB as the closest pairs, respectively. This genome-wide observation supports previous findings of discordant phylogenetic signals within *Aegilops* and *Triticum* genera (*6*, *43*, *45*). Some variation in genome relationships was found among chromosomes: On group 4 chromosomes, most gene trees supported BD as closest pairs, whereas group 5 chromosomes had similar numbers of AD and BD topologies (AD = BD > AB). Distribution of variation in phylogenetic signals across homeologous chromosomes can help to better understand the nature of the evolutionary processes underlying such phylogenetic incongruence. Under incomplete lineage sorting and stochastic coalescence, levels of phylogenetic incongruence will be correlated with recombination rates, whereas single introgression events and limited recombination are expected to generate local chromosome blocks of homogenous phylogenetic signals. We used the inferred gene orders from the GenomeZipper to test for nonrandom distribution of phylogenetic signals along chromosomes. We were unable to consistently identify block structures larger than would be expected by chance. However, it is possible that the limitations of the inferred gene order hamper the ability to detect such patterns.

## Gene expression

Our study did not reveal any pronounced bias in gene content, structure, or composition between the different wheat subgenomes. In paleopolyploid maize and soybean, transcriptional dominance of genes derived from one progenitor genome has been described (*64*–*66*). Previous analyses have shown that rapid initiation of differential expression of homeologous wheat genes occurs upon polyploidization with a predominantly additive mode (*13*, *67*). Sets of homeologous wheat genes with only one copy present in each of the subgenomes (triads) were used to test for differential expression at a genome-wide scale. Expression correlations were calculated for 6219 triads (18,657 genes) by using RNA-seq data from five organs (leaf, root, grain, spike, and stem) (Fig. 5A) (*25*). Whereas root-derived expression clustered separately, genes expressed in stem, leaves, grain, and spike clustered in a subgenome-specific manner. This indicates that the individual subgenomes exhibit a high degree of regulatory and transcriptional autonomy, with limited trans (inter-subgenome) regulation (*68*). At a global level, the overall pairwise expression correlation between subgenomes was very similar (Fig. 5B), and no evidence for genome-wide transcriptional dominance of an individual subgenome was observed.

By using hierarchical cluster analysis, we aggregated expressed genes into 13 distinct groups. These groups show predominant expression in particular organs (e.g., groups III and XIII in Fig. 5A) or in one of the subgenomes (e.g., groups II, IX, and X in Fig. 5A). Pairwise comparisons of individual expressed homeologous genes in the groups revealed abundant transcriptional dominance from specific subgenomes (Fig. 5B). Overall, 1333 (21%) of the homeologous gene triads showed an expression bias in one of the pairwise comparisons, and we detected a similar number of preferentially transcribed genes (378 to 393) in each subgenome (permutation test; $P < 0.05$). For the individual transcriptional groups, however, between 2% (groups I, IV, and V) and 20% (groups II and VI) of the genes were found to be transcriptionally dominant.

These patterns of gene expression across the three genomes contrast with patterns of gene expression reported in allopolyploid cotton (*69*, *70*); mesopolyploid *Brassica rapa* (*71*); synthetic allotetraploid *Arabidopsis* (*72*); and the paleopolyploid maize genome (*64*), where one of the genomes is more transcriptionally active than others. The apparent autonomy of the three wheat subgenomes may be explained by the relatively recent polyploidization. It may also be related to regulatory mechanisms that control the transcriptional interplay of homeologous genomes to balance expression of individual and groups of genes. While maintaining subgenome-specific expression profiles, a high degree of orchestration and functional partitioning between homeologous genes was also reported in grain development of bread wheat (*68*) and has been attributed to the rapid evolution of cis elements coupled to epigenetic mechanisms controlling gene expression (*68*, *73*, *74*).

## Gene family size variation

The relationship between genes important to wheat adaptation, disease resistance, and end-use functionality in hexaploid wheat and its diploid relatives was examined for signs of adaptive evolution. These analyses identified three distinct patterns: gene expansion, gene loss, or independent gene evolution that may or may not include expansion or loss. In some cases, such as the genes containing a NB-ARC domain characteristic of many plant disease-resistance genes (*75*), we observed an expansion within a single subgenome (Fig. 6A). Indeed, a substantial expansion in *Ae. tauschii*, compared with the other diploid species and the D genome of hexaploid wheat, is consistent with the rich reservoir of disease-resistance genes known in this species

(*17*). In genes coding for the cysteine-rich gliadin domain, a functional domain characteristic of storage proteins, we observed a similar number of genes in all diploid genomes (except *T. monococcum*) that is higher than the number of genes found in each of the three hexaploid wheat subgenomes (Fig. 6B). This may indicate that gene loss occurred in hexaploid wheat and that there is a trend for the gliadin gene family to maintain some homeostasis with a similar global number of genes in polyploid and diploid wheat. In other cases, the patterns observed suggested independent evolution of gene families within the different genomes and subgenomes of wheat. This was seen for genes associated with abiotic stress tolerance. For example, for genes encoding the Apetala2 (AP2) DNA binding domain, associated with drought, heat, salinity, and cold stress–tolerance responses, we observed fewer AP2 genes in the A and D genomes of Chinese Spring compared with the diploid relatives or the B subgenome (Fig. 6C). Likewise, genes coding for MYB transcription factors, which have also been involved in abiotic stress response in plants (*76*), were underrepresented in the A subgenome of hexaploid wheat and *T. monococcum*, whereas a higher frequency was observed in *Ae. tauschii* (*17*) and *T. urartu* (*16*) (Fig. 6D).

In contrast, there was no evidence of expansion or loss of genes underlying phenology, such as the vernalization (*Vrn1*) and photoperiod response regulator (*Ppd1*) genes that differentiate spring and winter growth habits and sensitivity to day length, respectively. Similar numbers of genes were found in the diploids and hexaploid subgenomes coding for the two functional domains of *Vrn1*, a MADS-box and K-box domain (*77*) (Fig. 6E), and for genes containing the response regulator domain and CCT motif typical of cereal *Ppd* genes (*78*) (Fig. 6F). We identified an additional copy of a *Vrn1*-like gene in the hexaploid Chinese Spring A and D genomes and *T. urartu* (*16*) when compared with the remaining diploid species. An additional copy of a *Ppd1*-like gene was also identified in the Chinese Spring B genome relative to *Ae. sharonesis* and *Ae. speltoides* (Fig. 6F). Although only small differences were observed, small increases in copy number variation of *Vrn-A1* (A genome) and *Ppd-B1* (B genome) have been associated with longer periods of vernalization to potentiate flowering and an early flowering day neutral phenotype, respectively (*79*). Thus, the relative distribution of such patterns in ontology of these two genes is likely to reflect important factors that have allowed wheat to adjust its flowering time to adapt to a range of environmental conditions.

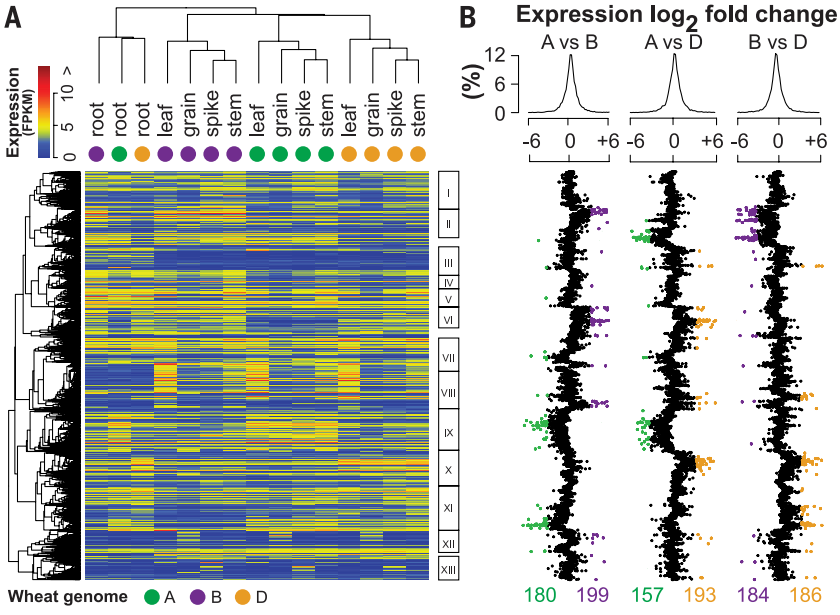



**Fig. 5. Subgenome transcriptional profiling for individual wheat tissues.** (**A**) Two-dimensional hierarchical cluster analysis of single-copy wheat homeologous gene expression (colors as in Fig. 2A) compared with organ-specific gene expression. (**B**) Analysis of $log_2$-fold changes in pairwise gene expression between homeologous genes (averaged across organs). Top graphs depict the distributions of $log_2$ fold changes. Dot plots show the fold changes for each triplet ordered as shown in the *y* axis in (A). Colored dots highlight homologs that show significant differential expression ($P < 0.05$). The numbers of differentially expressed triplets across all organs are shown at the bottom of the figure.

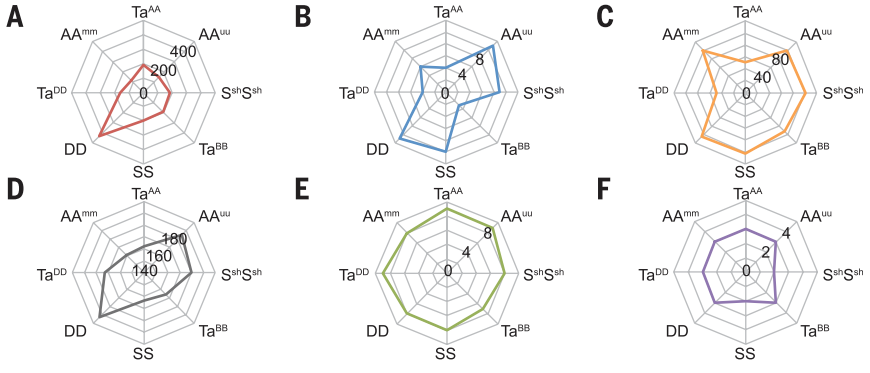



**Fig. 6. Sizes of selected gene families and protein domains among hexaploid wheat and diploid relatives.** (**A**) NB-ARC domain, (**B**) cysteine-rich gliadin domain, (**C**) AP2 domain, (**D**) MYB domain, (**E**) *Vrn1* (MADS-box/K-box domain), and (**F**) *Ppd* (photoperiod response regulator/CCT domain).

## Molecular markers

Wheat improvement relies in part on the use of molecular markers to improve selection efficiencies and to allow the precise transfer of genes and QTL between different genetic backgrounds. To enhance the CSS as a genomic resource for the wheat genetics and breeding community, we anchored all publicly available DNA markers that are routinely used for genetic mapping and marker-assisted breeding in wheat. Because the majority of these markers are anchored to phenotypic maps, anchoring them to the CSS allows immediate association of CSS to traits targeted by breeders. In addition, insertion site–based polymorphism (ISBP) and SNP markers identified from recent whole-genome shotgun and transcriptome sequencing (*19*) and genotyping by sequencing (GBS) tags identified by using DArTSeq (Diversity Arrays Technology, Bruce, Australia) technology were also anchored. In total, over 3.6 million marker loci were anchored to the CSS, including 1,347,669 marker loci and 2,310,988 SNPs (Table 5).

Most marker types showed a distribution gradient across subgenomes, with the highest number associated with the B genome chromosomes and the lowest with the D genome, reflecting the differences in the level of polymorphism in these subgenomes. The proportions of ISBPs, SNPs detected from cultivar sequencing and GBS tags localized to the D genome ranged between 9.3 and 12%, with the lowest numbers mapping to the group 4 chromosomes (Table 5). Two hundred and ninety-two of 1867 simple sequence repeat (SSR) loci were successfully anchored to the CSS survey sequence. This low number is not surprising, given that these loci derive from repetitive AT- and GC-rich sequences that may be collapsed or

**Table 5. Number and type of molecular markers mapped on individual chromosomes of the bread wheat genome.**

| | Bin mapped ESTs | EST-SSRs | Genomic SSRs | DArT Probes | Cereals DB | 90K iSelect SNPs (87) | DArT Seq | ISBPs | Genic SNPs | Intergenic SNPs | Σ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Queries | 18,771 | 2,926 | 1,867 | 7,552 | 7,228 | 81,987 | 29,375 | Derived from cultivar sequencing | | | - |
| Mapped queries | 16,876 | 2,435 | 282 | 5,228 | 5,136 | 80,820 | 18,515 | | | | |
| 1A | 1,325 | 156 | 8 | 414 | 479 | 13,093 | 1,371 | 68,074 | 13,980 | 127,663 | 226,563 |
| 2A | 1,614 | 257 | 28 | 356 | 544 | 17,502 | 1,378 | 84,440 | 18,349 | 148,204 | 272,672 |
| 3A | 1,136 | 75 | 14 | 252 | 302 | 12,172 | 1,008 | 44,740 | 10,770 | 94,975 | 165,444 |
| 4A | 1,766 | 266 | 27 | 331 | 357 | 14,043 | 1,530 | 39,483 | 10,367 | 86,543 | 154,713 |
| 5A | 1,189 | 155 | 46 | 256 | 343 | 13,099 | 893 | 62,193 | 12,624 | 115,085 | 205,883 |
| 6A | 1,150 | 132 | 63 | 418 | 421 | 12,072 | 1,127 | 60,169 | 15,884 | 110,850 | 202,286 |
| 7A | 1,240 | 146 | 120 | 321 | 326 | 13,168 | 1,474 | 71,597 | 15,516 | 154,748 | 258,656 |
| Σ A genome | 9,420 | 1,187 | 306 | 2,348 | 2,772 | 95,149 | 8,781 | 430,696 | 97,490 | 838,068 | 1,486,217 |
| 1B | 1,379 | 226 | 15 | 378 | 618 | 13,776 | 1,846 | 66,994 | 14,447 | 131,682 | 231,361 |
| 2B | 1,810 | 367 | 39 | 466 | 606 | 18,352 | 2,557 | 90,852 | 23,958 | 162,335 | 301,342 |
| 3B | 1,845 | 188 | 29 | 406 | 444 | 14,471 | 2,294 | 108,810 | 22,032 | 208,306 | 358,825 |
| 4B | 1,401 | 188 | 42 | 278 | 294 | 11,019 | 856 | 36,937 | 7,506 | 59,175 | 117,696 |
| 5B | 1,911 | 343 | 86 | 399 | 527 | 17,087 | 2,112 | 84,179 | 21,389 | 159,359 | 287,392 |
| 6B | 978 | 43 | 139 | 320 | 313 | 12,448 | 1,171 | 65,982 | 11,974 | 130,463 | 223,831 |
| 7B | 999 | 107 | 151 | 270 | 205 | 11,635 | 1,123 | 72,307 | 10,997 | 136,932 | 234,726 |
| Σ B genome | 10,323 | 1,462 | 501 | 2,517 | 3,007 | 98,788 | 11,959 | 526,061 | 112,303 | 988,252 | 1,755,173 |
| 1D | 1,165 | 149 | 13 | 378 | 380 | 12,093 | 660 | 17,366 | 5,004 | 36,457 | 73,665 |
| 2D | 1,309 | 199 | 22 | 414 | 331 | 16,978 | 609 | 19,532 | 6,745 | 34,967 | 81,106 |
| 3D | 854 | 104 | 14 | 428 | 151 | 11,699 | 420 | 10,920 | 1,403 | 18,078 | 44,071 |
| 4D | 1,221 | 239 | 27 | 245 | 196 | 10,198 | 307 | 10,097 | 1,108 | 13,249 | 36,887 |
| 5D | 1,584 | 408 | 78 | 400 | 289 | 13,308 | 488 | 13,629 | 3,582 | 22,957 | 56,723 |
| 6D | 1,132 | 91 | 135 | 289 | 240 | 10,504 | 417 | 12,042 | 3,609 | 23,341 | 51,800 |
| 7D | 1,461 | 230 | 139 | 862 | 243 | 12,826 | 767 | 18,174 | 3,969 | 34,344 | 73,015 |
| Σ D genome | 8,726 | 1,420 | 428 | 3,016 | 1,830 | 87,606 | 3,668 | 101,760 | 25,420 | 183,393 | 417,267 |
| Σ | 28,469 | 4,069 | 1,235 | 7,881 | 7,609 | 281,543 | 24,408 | 1,058,517 | 235,213 | 2,009,713 | 3,658,657 |

represented by uneven read coverage in Illumina sequences (*80*).

Well over 70 DNA markers are routinely deployed by breeders for agronomic, pest resistance, and end-use quality, and most are available in the public domain (http://maswheat.ucdavis.edu). Anchoring of these to the CSS would facilitate identification of SNP markers for development of high-density marker maps, as a resource of correlated markers, and to aid map-based cloning of genes underlying important traits. In total, we anchored 68 of these markers to 74 contigs in the CSS. The application of the CSS in marker improvement was demonstrated with the CAPS (cleaved amplified polymorphic sequence) marker *Usw47*, which is linked to *Cdu-B1*, a gene responsible for reduced grain cadmium content in tetraploid wheat (*81*, *82*). Although *Usw47* is routinely used in marker-assisted selection, it is not amenable to high-throughput genotyping. Alignment of the *Usw47* sequence against the CSS mapped it to contig 5BL-10759151. This and eight neighboring contigs in the GenomeZipper contained 33 SNP markers, of which 5 were polymorphic in a doubled haploid mapping population used previously to localize *Cdu-B1*. Of the five SNP markers, two co-segregated, and the remainder flanked the gene by a single recombination event. These SNP markers can be readily implemented now in a high-throughput fashion to select for reduced grain cadmium content within breeding programs.

## Conclusion

We present the ordered and structured draft sequence of the bread wheat genome as well as a comparison between eight related wheat genomes. We defined a gene catalog for each of the 21 bread wheat chromosomes and positioned more than 75,000 genes along the chromosomes by using a combination of high-density wheat SNP mapping and synteny to sequenced grass genomes. In contrast to other species (*83*), polyploidization events in wheat did not cause a "genome shock" with subsequent rapid genome changes or functional dominance of one subgenome over the others. Intraspecific comparative analyses revealed a dynamic wheat genome with a high level of plasticity and a changing gene repertoire shaped by gene losses and gene-family expansions in all wheat genomes and subgenomes, with only a few species-specific genes. Through interspecific comparisons, we observed a higher abundance of intrachromosomal gene duplications in wheat compared with other grass genomes, which may be a mechanism for functional adaptation and underlie the global success of wheat as a cultivated crop.

The detection, chromosomal assignment, and description of a large proportion of the gene complement of bread wheat and their positional assignment on chromosome arms is a major milestone in facilitating the isolation of genes underlying agronomically important traits, providing a reference for future integration into systems biology, and improving wheat breeding efficiency. Already, the resources developed in this work have been used to support the analysis of selected wheat chromosomes (*20*, *41*, *84*–*86*). Last, as demonstrated by the completion of the reference sequence for chr. 3B (*23*), this draft genome sequence and complementary resources will support the assembly and annotation of the physical map–based reference sequences for the 21 bread wheat chromosomes.

## REFERENCES AND NOTES


1. D. B. Lobell, W. Schlenker, J. Costa-Roberts, Climate trends and global crop production since 1980. *Science* **333**, 616–620 (2011). doi: 10.1126/science.1204531; pmid: 21551030
2. Food and Agriculture Organization (FAO) of the United Nations, FAO cereal supply and demand brief (2013); www.fao.org/worldfoodsituation/csdb/en/.
3. D. Tilman, K. G. Cassman, P. A. Matson, R. Naylor, S. Polasky, Agricultural sustainability and intensive production practices. *Nature* **418**, 671–677 (2002). doi: 10.1038/nature01014; pmid: 12167873
4. J. A. Foley *et al.*, Solutions for a cultivated planet. *Nature* **478**, 337–342 (2011). doi: 10.1038/nature10452; pmid: 21993620
5. Organisation for Economic Cooperation and Development (OECD)/FAO, *OECD-FAO Agricultural Outlook 2013* (OECD, Paris, 2013); doi: 10.1787/agr_outlook-2013-en.
6. G. Petersen, O. Seberg, M. Yde, K. Berthelsen, Phylogenetic relationships of *Triticum* and *Aegilops* and evidence for the

origin of the A, B, and D genomes of common wheat (*Triticum aestivum*). *Mol. Phylogenet. Evol.* **39**, 70–82 (2006). doi: 10.1016/j.ympev.2006.01.023; pmid: 16504543
7. M. Nesbitt, D. Samuel, "From staple crop to extinction? The archaeology and history of the hulled wheats," in *Hulled Wheat: Proceedings of the First International Workshop on Hulled Wheats*, S. Padulosi, K. Hammer, J. Heller, Eds. (International Plant Genetic Resources Institute, Rome, 1995), pp. 41–102.
8. E. Martinez-Perez, P. Shaw, G. Moore, The Ph1 locus is needed to ensure specific somatic and meiotic centromere association. *Nature* **411**, 204–207 (2001). doi: 10.1038/35075597; pmid: 11346798
9. T. Eilam *et al*., Genome size and genome evolution in diploid Triticeae species. *Genome* **50**, 1029–1037 (2007). doi: 10.1139/G07-083; pmid: 18059548
10. T. Wicker *et al*., Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives. *Plant Cell* **23**, 1706–1718 (2011). doi: 10.1105/tpc.111.086629; pmid: 21622801
11. K. Mochida, T. Yoshida, T. Sakurai, Y. Ogihara, K. Shinozaki, TriFLDB: A database of clustered full-length coding sequences from Triticeae with applications to comparative grass genomics. *Plant Physiol.* **150**, 1135–1146 (2009). doi: 10.1104/pp.109.138214; pmid: 19448038
12. A. N. Bernardo *et al*., Discovery and mapping of single feature polymorphisms in wheat using Affymetrix arrays. *BMC Genomics* **10**, 251 (2009). doi: 10.1186/1471-2164-10-251; pmid: 19480702
13. H. Chelaifa *et al*., Prevalence of gene expression additivity in genetically stable wheat allohexaploids. *New Phytol.* **197**, 730–736 (2013). doi: 10.1111/nph.12108; pmid: 23278496
14. T. E. Coram, M. L. Settles, M. Wang, X. Chen, Surveying expression level polymorphism and single-feature polymorphism in near-isogenic wheat lines differing for the Yr5 stripe rust resistance locus. *Theor. Appl. Genet.* **117**, 401–411 (2008). doi: 10.1007/s00122-008-0784-5; pmid: 18470504
15. L. L. Qi *et al*., A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. *Genetics* **168**, 701–712 (2004). doi: 10.1534/genetics.104.034868; pmid: 15514046
16. H. Q. Ling *et al*., Draft genome of the wheat A-genome progenitor *Triticum urartu*. *Nature* **496**, 87–90 (2013). doi: 10.1038/nature11997; pmid: 23535596
17. J. Jia *et al*., *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature* **496**, 91–95 (2013). doi: 10.1038/nature12028; pmid: 23535592
18. R. Brenchley *et al*., Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* **491**, 705–710 (2012). doi: 10.1038/nature11650; pmid: 23192148
19. A. M. Allen *et al*., Discovery and development of exome-based, co-dominant single nucleotide polymorphism markers in hexaploid wheat (*Triticum aestivum* L.). *Plant Biotechnol. J.* **11**, 279–295 (2013). doi: 10.1111/pbi.12009; pmid: 23279710
20. K. V. Krasileva *et al*., Separating homeologs by phasing in the tetraploid wheat transcriptome. *Genome Biol.* **14**, R66 (2013). doi: 10.1186/gb-2013-14-6-r66; pmid: 23800085
21. C. Saintenac, D. Jiang, S. Wang, E. Akhunov, Sequence-based mapping of the polyploid wheat genome. *G3* **3**, 1105–1114 (2013).
22. E. Sears, L. Sears, "The telocentric chromosomes of common wheat," in *Proceedings 5th International Wheat Genetics Symposium*, S. Ramanujam, Ed. (Indian Agricultural Research Institute, New Delhi, 1978) vol. 1, pp. 389–407.
23. F. Choulet *et al*., A reference sequence of wheat chromosome 3B reveals structural and functional compartmentalization. *Science* **345**, 1249721 (2014).
24. J. Šafář *et al*., Development of chromosome-specific BAC resources for genomics of bread wheat. *Cytogenet. Genome Res.* **129**, 211–223 (2010). doi: 10.1159/000313072; pmid: 20501977
25. Materials and methods are available as supporting materials on *Science* Online.
26. J. T. Simpson *et al*., ABySS: A parallel assembler for short read sequence data. *Genome Res.* **19**, 1117–1123 (2009). doi: 10.1101/gr.089532.108; pmid: 19251739
27. K. F. Mayer *et al*., A physical, genetic, and functional sequence assembly of the barley genome. *Nature* **491**, 711–716 (2012). pmid: 23075845
28. S. Kurtz, A. Narechania, J. C. Stein, D. Ware, A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics* **9**, 517 (2008). doi: 10.1186/1471-2164-9-517; pmid: 18976482
29. J. D. Hollister, B. S. Gaut, Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res.* **19**, 1419–1428 (2009). doi: 10.1101/gr.091678.109; pmid: 19478138
30. M. Kantar *et al*., Subgenomic analysis of microRNAs in polyploid wheat. *Funct. Integr. Genomics* **12**, 465–479 (2012). doi: 10.1007/s10142-012-0285-0; pmid: 22592659
31. S. J. Lucas, H. Budak, Sorting the wheat from the chaff: Identifying miRNAs in genomic survey sequences of *Triticum aestivum* chromosome 1AL. *PLOS ONE* **7**, e40859 (2012). doi: 10.1371/journal.pone.0040859; pmid: 22815845
32. G. M. Borchert *et al*., Comprehensive analysis of microRNA genomic loci identifies pervasive repetitive-element origins. *Mob. Genet. Elements* **1**, 8–17 (2011). doi: 10.4161/mge.1.1.15766; pmid: 22016841
33. International Brachypodium Initiative, Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**, 763–768 (2010). doi: 10.1038/nature08747; pmid: 20148030
34. International Rice Genome Sequencing Project, The map-based sequence of the rice genome. *Nature* **436**, 793–800 (2005). doi: 10.1038/nature03895; pmid: 16100779
35. A. H. Paterson *et al*., The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**, 551–556 (2009). doi: 10.1038/nature07723; pmid: 19189423
36. F. Choulet *et al*., Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. *Plant Cell* **22**, 1686–1701 (2010). doi: 10.1105/tpc.110.074187; pmid: 20581307
37. T. Lu *et al*., Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. *Genome Res.* **20**, 1238–1249 (2010). doi: 10.1101/gr.106120.110; pmid: 20627892
38. Y. Okazaki *et al*., Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563–573 (2002). doi: 10.1038/nature01266; pmid: 12466851
39. Y. Marquez, J. W. Brown, C. Simpson, A. Barta, M. Kalyna, Transcriptome survey reveals increased complexity of the alternative splicing landscape in *Arabidopsis*. *Genome Res.* **22**, 1184–1195 (2012). doi: 10.1101/gr.134106.111; pmid: 22391557
40. M. M. Martis *et al*., Reticulate evolution of the rye genome. *Plant Cell* **25**, 3685–3698 (2013). doi: 10.1105/tpc.113.114553; pmid: 24104565
41. P. Hernandez *et al*., Next-generation sequencing and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content. *Plant J.* **69**, 377–386 (2012). doi: 10.1111/j.1365-313X.2011.04808.x; pmid: 21974774
42. J. Ma *et al*., Sequence-based analysis of translocations and inversions in bread wheat (*Triticum aestivum* L.). *PLOS ONE* **8**, e79329 (2013). doi: 10.1371/journal.pone.0079329; pmid: 24260197
43. J. S. Escobar *et al*., Multigenic phylogeny and analysis of tree incongruences in Triticeae (Poaceae). *BMC Evol. Biol.* **11**, 181 (2011). doi: 10.1186/1471-2148-11-181; pmid: 21702931
44. P. Cíváň, Z. Ivaničová, T. A. Brown, Reticulated origin of domesticated emmer wheat supports a dynamic model for the emergence of agriculture in the fertile crescent. *PLOS ONE* **8**, e81955 (2013). doi: 10.1371/journal.pone.0081955; pmid: 24312385
45. T. Marcussen *et al*., Ancient hybridizations among the ancestral genomes of bread wheat. *Science* **345**, 1250092 (2014).
46. S. Griffiths *et al*., Molecular characterization of Ph1 as a major chromosome pairing locus in polyploid wheat. *Nature* **439**, 749–752 (2006). doi: 10.1038/nature04434; pmid: 16467840
47. K. F. X. Mayer *et al*., Gene content and virtual gene order of barley chromosome 1H. *Plant Physiol.* **151**, 496–505 (2009). doi: 10.1104/pp.109.142612; pmid: 19692534
48. G. Moore, K. M. Devos, Z. Wang, M. D. Gale, Cereal genome evolution. Grasses, line up and form a circle. *Curr. Biol.* **5**, 737–739 (1995). doi: 10.1016/S0960-9822(95)00148-5; pmid: 7583118
49. M. C. Luo *et al*., A 4-gigabase physical map unlocks the structure and evolution of the complex genome of *Aegilops tauschii*, the wheat D-genome progenitor. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 7940–7945 (2013). doi: 10.1073/pnas.1219082110; pmid: 23610408
50. M. Mascher *et al*., Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). *Plant J.* **76**, 718–727 (2013). doi: 10.1111/tpj.12319; pmid: 23998490
51. M. E. Sorrells *et al*., Reconstruction of the synthetic W7984 x Opata M85 wheat reference population. *Genome* **54**, 875–882 (2011). doi: 10.1139/g11-054; pmid: 21999208
52. J. Zhang, Evolution by gene duplication: An update. *Trends Ecol. Evol.* **18**, 292–298 (2003). doi: 10.1016/S0169-5347(03)00033-8
53. L. Li, C. J. Stoeckert Jr., D. S. Roos, OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003). doi: 10.1101/gr.1224503; pmid: 12952885
54. R. Koszul, S. Caburet, B. Dujon, G. Fischer, Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. *EMBO J.* **23**, 234–243 (2004). doi: 10.1038/sj.emboj.7600024; pmid: 14685272
55. J. L. Bennetzen *et al*., Reference genome sequence of the model plant *Setaria*. *Nat. Biotechnol.* **30**, 555–561 (2012). doi: 10.1038/nbt.2196; pmid: 22580951
56. P. S. Schnable *et al*., The B73 maize genome: Complexity, diversity, and dynamics. *Science* **326**, 1112–1115 (2009). doi: 10.1126/science.1178534; pmid: 19965430
57. T. Tanaka *et al*., The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucleic Acids Res.* **36**, D1028–D1033 (2008).pmid: 18089549
58. H. Ozkan, A. A. Levy, M. Feldman, Allopolyploidy-induced rapid genome evolution in the wheat (*Aegilops-Triticum*) group. *Plant Cell* **13**, 1735–1747 (2001). doi: 10.1105/tpc.13.8.1735; pmid: 11487689
59. R. J. Buggs *et al*., Rapid, repeated, and clustered loss of duplicate genes in allopolyploid plant populations of independent origin. *Curr. Biol.* **22**, 248–252 (2012). doi: 10.1016/j.cub.2011.12.027; pmid: 22264605
60. A. H. Paterson *et al*., Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature* **492**, 423–427 (2012). doi: 10.1038/nature11798; pmid: 23257886
61. R. Grantham, Amino acid difference formula to help explain protein evolution. *Science* **185**, 862–864 (1974). doi: 10.1126/science.185.4154.862; pmid: 4843792
62. J. Cao *et al*., Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat. Genet.* **43**, 956–963 (2011). doi: 10.1038/ng.911; pmid: 21874002
63. E. D. Akhunov *et al*., Comparative analysis of syntenic genes in grass genomes reveals accelerated rates of gene structure and coding sequence evolution in polyploid wheat. *Plant Physiol.* **161**, 252–265 (2013). doi: 10.1104/pp.112.205161; pmid: 23124323
64. J. C. Schnable, N. M. Springer, M. Freeling, Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 4069–4074 (2011). doi: 10.1073/pnas.1101368108; pmid: 21368132
65. R. A. Rapp, J. A. Udall, J. F. Wendel, Genomic expression dominance in allopolyploids. *BMC Biol.* **7**, 18 (2009). doi: 10.1186/1741-7007-7-18; pmid: 19409075
66. B. Chaudhary *et al*., Reciprocal silencing, transcriptional bias and functional divergence of homeologs in polyploid cotton (Gossypium). *Genetics* **182**, 503–517 (2009). doi: 10.1534/genetics.109.102608; pmid: 19363125
67. M. Pumphrey, J. Bai, D. Laudencia-Chingcuanco, O. Anderson, B. S. Gill, Nonadditive expression of homoeologous genes is established upon polyploidization in hexaploid wheat. *Genetics* **181**, 1147–1157 (2009). doi: 10.1534/genetics.108.096941; pmid: 19104075
68. M. Pfeifer *et al*., Genome interplay in the grain transcriptome of hexaploid bread wheat. *Science* **345**, 1250091 (2014).
69. M. J. Yoo, E. Szadkowski, J. F. Wendel, Homoeolog expression bias and expression level dominance in allopolyploid cotton. *Heredity* **110**, 171–180 (2013). doi: 10.1038/hdy.2012.94; pmid: 23169565
70. K. L. Adams, R. Cronn, R. Percifield, J. F. Wendel, Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 4649–4654 (2003). doi: 10.1073/pnas.0630618100; pmid: 12665616
71. F. Cheng *et al*., Biased gene fractionation and dominant gene expression among the subgenomes of *Brassica rapa*. *PLOS ONE* **7**, e36442 (2012). doi: 10.1371/journal.pone.0036442; pmid: 22567157
72. J. Wang *et al*., Stochastic and epigenetic changes of gene expression in *Arabidopsis* polyploids. *Genetics* **167**, 1961–1973 (2004). doi: 10.1534/genetics.104.027896; pmid: 15342533
73. Z. J. Chen, Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. *Annu. Rev. Plant Biol.* **58**, 377–406 (2007). doi: 10.1146/annurev.arplant.58.032806.103835; pmid: 17280525
74. K. L. Adams, Evolution of duplicate gene expression in polyploid and hybrid plants. *J. Hered.* **98**, 136–141 (2007). doi: 10.1093/jhered/esl061; pmid: 17208934
75. G. van Ooijen *et al*., Structure-function analysis of the NB-ARC domain of plant disease resistance proteins. *J. Exp. Bot.* **59**, 1383–1397 (2008). doi: 10.1093/jxb/ern045; pmid: 18390848
76. A. Katiyar *et al*., Genome-wide classification and expression analysis of MYB transcription factor families in rice and

*Arabidopsis*. *BMC Genomics* **13**, 544 (2012). doi: 10.1186/1471-2164-13-544; pmid: 23050870
77. L. Yan *et al*., Positional cloning of the wheat vernalization gene VRN1. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 6263–6268 (2003). doi: 10.1073/pnas.0937399100; pmid: 12730378
78. A. Turner, J. Beales, S. Faure, R. P. Dunford, D. A. Laurie, The pseudo-response regulator Ppd-H1 provides adaptation to photoperiod in barley. *Science* **310**, 1031–1034 (2005). doi: 10.1126/science.1117619; pmid: 16284181
79. A. Díaz, M. Zikhali, A. S. Turner, P. Isaac, D. A. Laurie, Copy number variation affecting the Photoperiod-B1 and Vernalization-A1 genes is associated with altered flowering time in wheat (*Triticum aestivum*). *PLOS ONE* **7**, e33234 (2012). doi: 10.1371/journal.pone.0033234; pmid: 22457747
80. S. O. Oyola *et al*., Optimizing Illumina next-generation sequencing library preparation for extremely AT-biased genomes. *BMC Genomics* **13**, 1 (2012). doi: 10.1186/1471-2164-13-1; pmid: 22214261
81. R. E. Knox *et al*., Chromosomal location of the cadmium uptake gene (*Cdu1*) in durum wheat. *Genome* **52**, 741–747 (2009). doi: 10.1139/G09-042; pmid: 19935921
82. K. Wiebe *et al*., Targeted mapping of Cdu1, a major locus regulating grain cadmium concentration in durum wheat (*Triticum turgidum* L. var durum). *Theor. Appl. Genet.* **121**, 1047–1058 (2010). doi: 10.1007/s00122-010-1370-1; pmid: 20559817
83. L. Comai, The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* **6**, 836–846 (2005). doi: 10.1038/nrg1711; pmid: 16304599
84. P. J. Berkman *et al*., Sequencing and assembly of low copy and genic regions of isolated *Triticum aestivum* chromosome arm 7DS. *Plant Biotechnol. J.* **9**, 768–775 (2011). doi: 10.1111/j.1467-7652.2010.00587.x; pmid: 21356002
85. P. J. Berkman *et al*., Sequencing wheat chromosome arm 7BS delimits the 7BS/4AL translocation and reveals homoeologous gene conservation. *Theor. Appl. Genet.* **124**, 423–432 (2012). doi: 10.1007/s00122-011-1717-2; pmid: 22001910
86. T. Tanaka *et al*., Next-generation survey sequencing and the molecular organization of wheat chromosome 6B. *DNA Res.* **21**, 103–114 (2013). pmid: 24086083
87. S. Wang *et al*., Characterization of polyploid wheat genomic diversity using a high-density 90, 000 single nucleotide polymorphism array. *Plant Biotechnol. J.* (2014). doi: 10.1111/pbi.12183; pmid: 24646323

## ACKNOWLEDGMENTS

The authors would like to thank Graminor AS; Biogemma; Institut National de la Recherche Agronomique (INRA); International Center for Agricultural Research in the Dry Areas; Department of Biotechnology, Ministry of Science and Technology, Government of India (chr. 2A; grant no. BT/IWGSC/03/TF/2008); and the Biotechnology and Biological Sciences Research Council (BBSRC UK) for funding the chromosome sequencing at the Genome Analysis Centre. Chromosome sequencing at other centers was funded by the following: chr. 3A—U.S. Department of Agriculture Agriculture and Food Research Initiative (USDA AFRI) Triticeae-CAP (2011-68002-30029) and the Kansas Wheat Commission; chr. 3B—grants from the French National Research Agency (ANR-09- GENM-025 3BSEQ) and France Agrimer; chr. 6B—grants from the Ministry of Agriculture, Forestry and Fisheries of Japan "Genomics for agricultural innovation KGS-1003,1004", "Genomics based technology for agricultural improvement, NGB- 1003," and Nisshin Flour Milling Incorporated; chr. 6D and *Triticum durum* cv. Strongfield—grants from Genome Canada, Genome Prairie, University of Saskatchewan Ministry of Agriculture, Western Grains Research Foundation; chr. 7B—grant no. 199387 from the Norwegian Research Council and from Graminor AS; chr. 7A and 7D sequence reads were provided by D.E.. Chromosome flow sorting and DNA preparation was supported through grants P501/12/G090 and P501/12/2554 from the Czech Science foundation. Chromosome sequence assembly was supported by the BBSRC (UK). K.F.X.M. acknowledges grants from the German Ministry for Education and Research (BMBF) Plant2030, TRITEX, Deutsche Forschungsgemeinschaft (DFG) SFB 924, and EC Transplant. K.E. and J.R. are supported by sponsors of the IWGSC, which include Arcadia Biosciences, Australian Centre for Plant Functional Genomics, Biogemma, Bayer CropScience, Commonwealth Science and Industrial Research Organisation, Centro Internacional de Mejoramiento de Maíz y Trigo, Céréales Vallée, Dow AgroSciences, Dupont, Evogene, Florimond Desprez, Grains Research and Development Corporation, Graminor, Heartland Plant Innovation, INRA, KWS, Kansas Wheat Commission, Limagrain, Monsanto, RAGT, and Syngenta. N.G. is supported by European Commission Marie Curie Actions (FP7-MC-IIF-Noncollinear Genes). T.W. is supported by the Swiss National Foundation and P.F., M.C., A.M.S., and L.C. are supported by the Italian Ministry of Agriculture special project "MAPPA-5A." H.B. acknowledges funding from Sabanci University and the Scientific and Technological Research Council of Turkey. B.W. and B.S. were funded by the Gatsby Charitable Foundation and the BBSRC (UK) Grant BB/J003166/1. R.W. is a Trustee Director of TGAC, Norwich, UK, and A.K. is a shareholder of Diversity Arrays Technology Pty Ltd. The POPSeq analysis carried out by the U.S. Department of Energy Joint Genome Institute was supported by the Office of Science of the U.S. Department of Energy under contract no. DE-AC02-05CH11231. Additional support for the work was funded from the Triticeae-CAP, USDA AFRI (2011-68002-30029) to G.J.M.; the Scottish Government Rural and Environment Science and Analytical Services Division Research Programme to R.W.; and the German Ministry of Research and Education (BMBF TRITEX 0315954) to N.S. Sequence reads and assembled sequences are available at European Molecular Biology Laboratory/GenBank/DNA Data Bank of Japan short read archives and sequence repositories, respectively (PRJEB3955—whole-genome sequences of *T. aestivum* 'Chinese Spring,' *T. urartu*, *Ae. speltoides*, *Ae. tauschii*, *T. turgidum*; SRP004490.3—whole-genome sequencing of *T. monococcum*; SRP004490—whole-genome sequencing of *Ae. tauschii*; PRJEB4849—whole-genome sequences of *Ae. sharonensis*; PRJEB4750—*T. aestivum* RNA-seq data; SRP037990—*T. aestivum* SynOpDH mapping population; SRP037781—*T. aestivum* synthetic opata M85; SRP037994—*T. aestivum* synthetic W7984). All data can be accessed via the IWGSC repository at Unité de Rercherche Génomique Info: http://wheat-urgi.versailles.inra.fr/Seq-Repository/.

**The International Wheat Genome Sequencing Consortium (IWGSC)**
Authorship of this paper should be cited as "International Wheat Genome Sequencing Consortium." Participants are arranged by working group. Corresponding authors (*), major contributors (†), and equally contributing authors (‡) are indicated.
**Principal Investigators:** Klaus F. X. Mayer[1]* (k.mayer@helmholtz-muenchen.de), Jane Rogers[2]* (janerogersh@gmail.com), Jaroslav Doležel[3]* (dolezel@ueb.cas.cz), Curtis Pozniak[4]* (curtis.pozniak@usask.ca), Kellye Eversole[2]* (eversole@eversoleassociates.com), Catherine Feuillet[5]* (catherine.feuillet@bayer.com)
**Provision of seed material for ditelosomic wheat lines:** Bikram Gill,[6] Bernd Friebe,[6] Adam J. Lukaszewski,[7] Pierre Sourdille,[14] Takashi R Endo[8]
**Chromosome sorting and DNA preparation:** Jaroslav Doležel,[3]† Marie Kubaláková,[3] Jarmila Číhalíková,[3] Zdeňka Dubská,[3] Jan Vrána,[3] Romana Šperková,[3] Hana Šimková[3]
**DNA sequencing:** Jane Rogers,[2]† Melanie Febrer,[9] Leah Clissold,[10] Kirsten McLay,[10] Kuldeep Singh,[11] Parveen Chhuneja,[11] Nagendra K. Singh,[12] Jitendra Khurana,[13] Eduard Akhunov,[6] Frédéric Choulet,[14] Pierre Sourdille,[14] Catherine Feuillet,[5] Adriana Alberti,[15] Valérie Barbe,[15] Patrick Wincker,[15] Hiroyuki Kanamori,[16] Fuminori Kobayashi,[16] Takeshi Itoh,[16] Takashi Matsumoto,[16] Hiroaki Sakai,[16] Tsuyoshi Tanaka,[16] Jianzhong Wu,[16] Yasunari Ogihara,[17] Hirokazu Handa,[16] Curtis Pozniak,[4] P. Ron Maclachlan,[4] Andrew Sharpe,[18] Darrin Klassen,[18] David Edwards,[19] Jacqueline Batley,[19] Odd-Arne Olsen,[20,21] Simen Rød Sandve,[20] Sigbjørn Lien,[37] Burkhard Steuernagel,[22] Brande Wulff[22]
**DNA sequence assembly:** Mario Caccamo,[10]† Sarah Ayling,[10] Ricardo H. Ramirez-Gonzalez,[10] Bernardo J. Clavijo,[10] Burkhard Steuernagel,[22] Jonathan Wright[10]
**Gene annotation:** Matthias Pfeifer,[1] Manuel Spannagl,[1] Klaus F. X. Mayer[1]†
**Genome Zipping:** Mihaela M. Martis,[1] Eduard Akhunov,[6] Frédéric Choulet,[14] Klaus F. X. Mayer[1]†
**POPSEQ analysis:** Martin Mascher,[23] Jarrod Chapman,[24] Jesse A. Poland,[25] Uwe Scholz,[23] Kerrie Barry,[24] Robbie Waugh,[26] Daniel S. Rokhsar,[24] Gary J. Muehlbauer,[27] Nils Stein[28]
**Repetitive DNA analysis:** Heidrun Gundlach,[1] Matthias Zytnicki,[29] Véronique Jamilloux,[29] Hadi Quesneville,[29] Thomas Wicker,[30] Klaus F. X. Mayer[1]
**miRNAs:** Primetta Faccioli,[31]‡ Moreno Colaiacovo,[31]‡ Matthias Pfeifer,[1]‡ Antonio Michele Stanca,[31] Hikmet Budak,[32] Luigi Cattivelli[31]†
**Genome structure and duplications:** Natasha Glover,[14] Mihaela M. Martis,[1] Frédéric Choulet,[14] Catherine Feuillet,[5] Klaus F. X. Mayer[1]
**Transcriptome sequencing and expression analysis:** Matthias Pfeifer,[1] Lise Pingault,[14] Klaus F. X. Mayer,[1]† Etienne Paux[14]†
**Gene family analysis:** Manuel Spannagl,[1] Sapna Sharma,[1] Klaus F. X. Mayer,[1]† Curtis Pozniak[4]†
**Proteogenomics analysis:** Rudi Appels,[33]† Matthew Bellgard,[33] Brett Chapman,[33] Matthias Pfeifer[1]
**Comparative analysis of diploid, tetraploid and hexaploid wheat:** Matthias Pfeifer,[1] Simen Rød Sandve,[20] Thomas Nussbaumer,[1] Kai Christian Bader,[1] Frédéric Choulet,[14] Catherine Feuillet,[5] Klaus F. X. Mayer[1]†
**Development and mapping of marker sets:** Eduard Akhunov,[6] Etienne Paux,[14] Hélène Rimbert,[36] Shichen Wang,[6] Jesse A. Poland,[25] Ron Knox,[34] Andrzej Kilian,[35] Curtis Pozniak[4]†
**Sequence repository:** Michael Alaux,[29]† Françoise Alfama,[29] Loïc Couderc,[29] Véronique Jamilloux,[29] Nicolas Guilhot,[14] Claire Viseux,[29] Mikaël Loaec,[29] Hadi Quesneville[29]
**Study design:** Jane Rogers,[2] Jaroslav Doležel,[3] Kellye Eversole,[2] Catherine Feuillet,[5] Beat Keller,[30] Klaus F. X. Mayer,[1] Odd-Arne Olsen,[20,21] Sebastien Praud[36]

[1]Plant Genome and Systems Biology, Helmholtz Zentrum Munich, Ingolstädter Landstrasse 1, 85764 Neuherberg, Germany. [2]IWGSC, Eversole Associates, 5207 Wyoming Road, Bethesda, MD 20816, USA. [3]Institute of Experimental Botany, Center of Plant Structural and Functional Genomics, Šlechtitelů 31, 783 71 Olomouc, Czech Republic. [4]Crop Development Centre, Department of Plant Sciences, College of Agriculture and Bioresources, University of Saskatchewan, 51 Campus Drive, Saskatoon, SK, Canada. [5]Bayer Crop Science, 3500 Paramount Parkway, Morrisville, NC 27560, USA. [6]Kansas State University, Department of Plant Pathology, Manhattan, KS 66506–5502, USA. [7]College of Natural and Agricultural Sciences, Botany and Plant Sciences, University of California, Riverside, CA 92521, USA. [8]Laboratory of Plant Genetics, Graduate School of Agriculture, Kyoto University, Kyoto 606-8502, Japan. [9]Genomic Sequencing Unit, University of Dundee, Dow Street, Dundee DD1 5EH, UK. [10]Genome Analysis Centre, Norwich Research Park, Norwich, NR4 7UH, UK. [11]School of Agricultural Biotechnology, Punjab Agricultural University, Ludhiana 141 004, India. [12]National Research Centre on Plant Biotechnology, Indian Agricultural Research Institute, New Delhi 110 012, India. [13]Interdisciplinary Centre for Plant Genomics and Department of Plant Molecular Biology, University of Delhi, South Campus, New Delhi 110 021, India. [14]INRA–University Blaise Pascal UMR1095 Genetics, Diversity and Ecophysiology of Cereals, 5 chemin de Beaulieu, 63039 Clermont-Ferrand, France. [15]Commissariat à l'Energie Atomique Genoscope, Centre National de Séquençage, 2 rue Gaston Crémieux, CP5706, 91057 Evry, France. [16]Plant Genome Research Unit, National Institute of Agrobiological Sciences, 2-1-2, Kan-non-dai, Tsukuba 305-8602, Japan. [17]Kihara Institute for Biological Research, Yokohama City University, Maioka-cho 641-12, Totsuka-ku, 244-0813 Yokohama, Japan. [18]National Research Council Canada, 110 Gymnasium Place, Saskatoon, SK, S7N 0W9, Canada. [19]Australian Centre for Plant Functional Genomics, School of Agriculture and Food Sciences, University of Queensland, St. Lucia, QLD 4072, Australia, and School of Plant Biology, University of Western Australia, WA 6009, Australia. [20]Department of Plant Sciences, Center for Integrative Genetics (CIGENE), Norwegian University of Life Sciences, 1432 Ås, Norway. [21]Department of Natural Science and Technology, Hedmark University College, N-2318, Norway. [22]Sainsbury Laboratory, Norwich Research Park, Norwich, NR4 7UH, UK. [23]Bioinformatics and Information Technology, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), D-06466 Seeland OT Gatersleben, Germany. [24]U.S. Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA. [25]USDA-ARS Hard Winter Wheat Genetics Research Unit and Department of Agronomy, Kansas State University, Manhattan, KS 66506-5502, USA. [26]James Hutton Institute, Invergowrie, Dundee DD2 5DA, UK. [27]Department of Agronomy and Plant Genetics, Department of Plant Biology, University of Minnesota, St. Paul, MN 55108, USA. [28]Genome Diversity, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), D-06466 Seeland OT Gatersleben, Germany. [29]INRA, UR1164 URGI–Research Unit in Genomics-Info, INRA de Versailles, Route de Saint-Cyr, Versailles, 78026, France. [30]Institute of Plant Biology, University of Zurich, Zollikerstrasse 107, CH-8008 Zurich, Switzerland. [31]Consiglio per la Ricerca e la sperimentazione in Agricoltura–Genomics Research Centre, via San Protaso 302, I-29017 Fiorenzuola d'Arda, Italy. [32]Sabanci University Biological Sciences and Bioengineering Program, 34956 Istanbul, Turkey. [33]Centre for Comparative Genomics, Murdoch University, Perth, WA 6150, Australia. [34]Semiarid Prairie Agricultural Research Centre, Post Office Box 1030, Swift Current, Saskatchewan S9H 3X2, Canada. [35]Diversity Arrays Technology Pty Limited, 1 Wilf Crane Crescent, Yarralumla ACT2600, Australia. [36]Biogemma, Centre de Recherche de Chappes, Route d'Ennezat, 63720 Chappes, France. [37]Department of Animal and Aquicultural Sciences, CIGENE, Norwegian University of Life Sciences, Arboretvelen 6, 1432 Ås, Norway.

## Supplementary Materials

www.sciencemag.org/content/345/6194/1251788/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S60
Tables S1 to S48
References (*88*–*160*)

5 February 2014; accepted 2 June 2014
10.1126/science.1251788

**Annexe 3 :** Daron, J., Glover, N., **Pingault, L**., Theil, S., Jamilloux, V., Paux, E., … Choulet, F. (s. d.). Organization and Evolution of Transposable Elements along the Wheat Chromosome 3B.

# Title: Organization and Evolution of Transposable Elements along the Wheat Chromosome 3B

**Authors:** Josquin Daron[1,2], Natasha Glover[1,2], Lise Pingault[1,2], Sébastien Theil[1,2], Véronique Jamilloux[3], Etienne Paux[1,2], Valérie Barbe[4], Sophie Mangenot[4],Adriana Alberti[4], Patrick Wincker[4,5,6], Hadi Quesneville[3], Catherine Feuillet[1,2], Frédéric Choulet[1,2*]

**Affiliations:**

[1]INRA UMR1095 Genetics, Diversity and Ecophysiology of Cereals, 5 chemin de Beaulieu, 63039 Clermont-Ferrand, France
[2]University Blaise Pascal UMR1095 Genetics, Diversity and Ecophysiology of Cereals, 5 chemin de Beaulieu, 63039 Clermont-Ferrand, France
[3]INRA-URGI, Centre de Versailles, Route de Saint Cyr, 78026 Versailles, France
[4]CEA/DSV/IG/Genoscope, 2 rue Gaston Cremieux, 91000 Evry, France
[5]CNRS UMR 8030, 2 rue Gaston Crémieux 91000 Evry,France
[6]Université d'Evry, CP5706 Evry, France

*Corresponding author

**Email addresses:**

JD: josquin.daron@clermont.inra.fr
NG: natasha-marie.glover@clermont.inra.fr
LP: lise.pingault@gmail.com
ST: sebastien.theil@bordeaux.inra.fr
VJ: veronique.jamilloux@versailles.inra.fr
EP: etienne.paux@clermont.inra.fr
VB: vbarbe@genoscope.cns.fr
SM: smangenot@genoscope.cns.fr
AA: aalberti@genoscope.cns.fr
PW: pwincker@genoscope.cns.fr
HQ: hadi.quesneville@versailles.inra.fr
CF: catherine.feuillet@bayer.com
FC: frederic.choulet@clermont.inra.fr

## Abstract

### Background

The bread wheat genome has massively expanded (17 Gb) through the proliferation of transposable elements and two recent rounds of polyploidization. The assembly of a 774 Mb reference sequence of wheat chromosome 3B provided us with the opportunity to explore the impact of transposable elements (TEs) on the complex wheat genome structure and evolution at a resolution and scale never reached so far.

### Results

We developed an automated workflow (called CLARI-TE) for TE modeling in complex genomes. We delineated precisely 56,488 intact and 196,391 fragmented TEs along the 3B pseudomolecule, accounting for 85% of the sequence, and reconstructed 30,199 nested insertions. TEs are mainly silent since ca. 1 My and the actual 3B chromosome was shaped by a succession of bursts that have occurred between 3 to 1 Mya. Accelerated TE elimination in the high-recombination distal regions appeared as a driving force towards the chromosome partitioning. *CACTAs*overrepresented in the high-recombination distal regionswere found to be significantly associated with recently duplicated genes. In addition, we identified 140 *CACTA*-mediated gene capture events with 17 genes potentially created by exon shuffling and show that 19 captured genes are transcribed and under selection pressure, suggesting the important role of *CACTAs* in the recent wheat adaptation.

### Conclusion

Accurate TE modeling allowed us to decipher the dynamics of TEs in a highly complex and polyploidgenome. It provides novel insights into chromosome partitioning and highlights the role of *CACTA* transposons in the high level of gene duplication in wheat.

### Keywords

Transposable elements, wheat, annotation, genome evolution, gene duplication, CLARI-TE, *CACTA*

## Background

First discovered in maize [1], transposable elements (TEs) are ubiquitous components of almost every eukaryotic genome investigated so far and their impact on genome structure and evolution is now well established (for review [2]). Two classes of TEs have been defined: Class Ior retrotransposons that use the element-encoded mRNA as a transposition intermediate, and Class II or DNA transposons that excise from their insertion site and transpose through a DNA intermediate.Among sequenced plant genomes, the TE abundance ranges from 20% in *Arabidopsis thaliana*[3] to 85% in maize (*Zea mays*) [4]. Genome expansion is mainly mediated by the activity of class I elements, while the content of DNA transposons is much more constant [5-7]. Furthermore, TEsare non-randomly distributed along the genome due to insertions[8, 9]anddeletions[7, 10]that trigger genome expansion/contraction. Hence, in rice (*Oryza sativa*) [11], sorghum (*Sorghum bicolor*) [12] and maize, long terminal repeat retrotransposons (LTR-RTs) accumulate preferentially in heterochromatin such as centromeric regions, and are less abundant in the recombinogenic distal parts of chromosomes. The molecular mechanism responsible of the targeted integration of TE in heterochromatin has begun to be understood[13]. A key component of the TE integration complex is a chromodomain present at the C-terminal part of the element-encoded integrase in *gypsy*elements [14] that drives preferential insertion by targeting specific chromatin modifications [15]. Deletion of TEscan be attested by the identification of solo LTRs and truncated elements [6]. Presumably, solo LTRs are formed by unequal intra-chromosomal homologous recombination between two LTRs of an intact element [16]. In contrast, truncated elements are thought to be the outcome of illegitimate (nonhomologous) recombination. In rice, 190Mb of LTR-RTs DNA have been removed recently by these two processes, leaving a current genome of ~400 Mb that contains <100 Mb of detectable LTR-RTs [17]. In sorghum and rice, DNA transposons are localized essentially in the telomeric regions of the chromosome[11, 12]. However, no evidence for a direct relationship between genetic recombination rate and DNA transposon abundance has been provided so far in plant genomes [18]whereas it was observed in *Caenorhabditis elegans*[19]. A key aspect of DNA transposons is their interaction with host genes [20]. For instance, they were shown to be involved in the creation of new genes through “exon shuffling”. In maize, 60% of the 20,000 *Helitrons* contain captured gene fragments [21, 22]. Similarly, in rice, 2809 Pack-Mutator-like elements (Pack-MULEs) containing host gene fragments were identified [23].

Identification of transposable elements in large and complex genomes is a daunting task, even in high quality assembled genomes. As an example, since 2001, in the rice genome, the number of MITEs discovered has steadily increased over time from 6641 to 179,415 [24-28]. A high-quality TE prediction and annotation is essential to prevent mis-annotation of functional genes and to

understand the biology of genomes [29].Different strategies have been developed, including similarity search against databanks of known TE sequences, *de novo* repeat detection, k-mer based counting, and structural motif detection (for review [30]). Despite the development of many dedicated bioinformatic tools, precise TE modeling in complex (>1 Gb) genomes, such as in wheat or maize, is a tour de force. During the maize genome sequencing project (~2 Gb), TEs were predicted with a variety of approaches: LTR-RTs and *Helitrons* were identified by structural criteria, while the rest of the TEs were annotated by similarity with a library built by *de novo* detection [4]. One of the main limitationsto identify TEs, and especially large TEs (>5 kb), was the fragmentation of the assembly with many gaps and a short median contig size (~7 kb).

The wheat genome is large and highly complex (17 Gb, allohexaploid 2n=6x=42 with 3 closely related subgenomes AABBDD). Previous small-scale analyses revealed that the wheat genome is composed of about 80% TE-derived sequences, mainly nested into each other and with a few families representing 50% of the TE fraction [31].Early, manually curated-TE modeling has been performed on selected BAC sequences during map-based cloning projects or using plasmid or BAC end sequences [32-36]. At a larger scale, analyses of 18 Mb of long BAC contig sequences spread along the 3B chromosome [31]led to precisely delineate 3222 TEs. Together with the public *Triticeae* REPeat sequence database (TREP) (http://wheat.pw.usda.gov/ITMI/Repeats/), this provided a representative high quality reference library of wheat TEs. Beyond the identification of TEs, the reconstruction of the nested insertion pattern isa computational challengewhich requiresfine tuning of dedicated algorithms and knowledge of TE structure and evolution.

Recently, a number of genome scale sequences were produced from different wheat genome. Whole genome shotgun sequencing, using short read sequencing technologies have been produced from bread wheat and from diploid species related to the homoeologous A and D genomes [37-39]. While these have been useful to characterize the gene space, the assemblies were highly fragmented and therefore had only limited value to study TEs. Moreover, they do not provide sufficient sequence contiguity to assemble pseudomolecules precluding the analysis of any TE feature distribution along a chromosome. To obtain a reference genome sequence of bread wheat, the International Wheat Genome Sequencing Consortium (IWGSC; www.wheatgenome.org) has established an approach based on flow sorting of individual chromosomes and the construction and sequencing of chromosome-specific physical maps. Recently, we produced a pseudomolecule for the largest wheat chromosome (3B) which represents 774 Mb, carrying 7,264 genes and 85% of TE-derived sequences[40]. Gene annotation and comparative analyses indicated that chromosome 3B and the wheat genome in general [41]carries a higher number of genes than related grass species. Moreover, the results showed that about35% of the 3B genes shares similarity with genes

located on non-orthologous chromosomes in other grasses[40]. These 'nonsyntenic' genes likely originate from interchromosomal duplications triggered by diverse mechanisms such as double strand break repair or TE mobilization.

In this study, a strategy dedicated to TE-modeling in complex genome was developed to decipher the complex organization of TEs along a wheat chromosome. Analyses of the distribution of the abundance, diversity and dynamics of TEs revealed a striking partitioning of the chromosome as observed for other features on chromosome 3B [40]. In addition, we observed a massive amplification of *CACTA* DNA transposon compared to other related grass species with a significant association between some *CACTA* families and recently duplicated genes, suggesting a role for*CACTA* transposons in gene duplications, gene capture, and genome plasticity in wheat.

## Results

### *An improved procedure for predicting TE models and their nested pattern in a complex genome*

Predicting TE features in complex genomes where repeated elements represent more than 80% of the sequence remains a computational challenge, and obtaining a high quality annotation still requires manual curation. Typically, TE prediction is performed by similarity search with a set of known TE sequences.A major prerequisite to achieve a high quality annotation is the availability of a curated TE reference library. This is essential to identify most transposons via similarity search-based approaches and to restrict the *de novo* detectionof repeats to the unassigned portion of the genome sequence. In wheat, two curated libraries dedicated to *Triticeae* (wheat, barley and rye) transposable elements are available: the *Triticeae* REPeat sequence database (TREP), which contains 1717 TEs representing 323 families, and an additional set of 3212 TEs manually annotated in a previous pilot study of chromosome 3B[31]. For most of them, the borders of the mobile element have been defined precisely and their completeness i.e., “complete element” *versus* “fragmented element”,is also available. However, there are incongruences in the classification of the TEsmost of them TEs were assigned a family name based on their best BLAST hit, which can lead to an over-estimation of the family numbers[42].

To ensure a proper TE modeling of the 774 Mb of sequence from chromosome 3B, we first built a classified TE library (ClariTeRep; see Material and Methods) using the 3050 complete TE sequences from these two libraries. In total, 525 families, comprising1 to 266 copies each, were clustered based on their sequence similarity. Among the families described in TREP, 40% (277/700) have a 1-to-1 relationship with the ClariTeRep classification while the others were mostly aggregated into a single family and, sometimes, split into different families. More than 80% of the TREP families comprise 1 to 2 TE copies only *versus* 57% for ClariTeRep, confirming that the similarity-based classification increases the number of families compared to the clustering-based approach [42].

In the second step, we automated two things: (i) the correction of the TE prediction’s over-fragmentation, i.e. the fact that a TE is not detected as a single feature but rather split into several neighboring ones, and (ii) the reconstruction of nested TE insertions.To this aim, we developed a program called CLARI-TE, which allowsmergingneighboring predictions that belong to the same family (see Material and Methods).Then, nested clusters were automatically reconstructed by joining remotepredictions belonging to a single element.We estimated the accuracy of the automated annotation by analyzing a manually annotated sequence of ~1 Mb (scaffold v443_0137)containing

196 TEs and 47 nested insertions, that was manually annotated for this purpose. We compared the annotation produced by RepeatMasker (using ClariTeRep) to its improvement when using CLARI-TE and other annotation pipeline, TEannot (part of the REPET package [43]) (Table 1). At the nucleotide level, RepeatMasker correctly assigned ~90% of nucleotides which needed belong to TEs, showing that our TE library is comprehensive enough to detect the vast majority of TEs in a wheat genomic sequence.However, the 196 TEs were predicted as 590 separated features with RepeatMasker, illustrating the over-fragmentation problem. The improvement was significant when using both TEannot and CLARI-TE with the detection of 345 and 289 features, respectively (Table 1).CLARI-TEwas more accurate than TEannot in predicting the correct borders of the TEs (sensibility: 66% vs. 45%, respectively) and inlimiting the number of false positives (specificity: 52% vs. 27%, respectively).In addition, CLARI-TE was able to perform the reconstruction of nested clusters much more accurately than TEannot (Table 1). The accuracy of nested insertion mining was in fact highly dependent on the type of TEs: sensitivity and specificity were much higher for *gypsy* and *copia*(53% and 68%, respectively) than for *CACTAs* (13% and 40%, respectively). This suggests that *CACTAs* exhibit a higher level of sequence variability and that *CACTA*-typical short tandem repeats limit our ability to identify them through fully automated procedures.

### *TE content and distribution along chromosome 3B*

We used CLARI-TE to predict TE-models along the 833 Mb of chromosome 3B, sequence corresponding to the 3B pseudomolecule [774 Mb] and unanchored scaffolds [59 Mb]), in order to study the organization and dynamics of TEs along a wheat chromosome. In total, 523,233 RepeatMasker-detected features were combined with CLARI-TE to obtain a final set of 56,488 complete TEs (*i.e.* with 2 borders corresponding to the borders of a reference element) and 196,391 truncated or partially assembled elements. Thus, using CLARI-TE we are able to reduce the fragmentation of the annotation by a 2-fold factor and detect the largest set of full-length TEs and nested clusters ever observed on a plant chromosome so far. The number of large TEs (>5 kb) was doubled (from 16,988 to 30,894)when using CLARI-TE. Moreover, 30,199 nested insertions were reconstructed comprising up to 8 layers and 320 clusters larger than 100 kb (and up to 301 kb). Overall, TEs represent 85% of the 833 Mb of chromosome 3B scaffolds, including the *de novo* identification of 3% of previously uncharacterized TEs[40].We applied the same approach to annotate the draft genome sequences of the A- and D-related diploid progenitors: *Triticum urartu* (~3 Gb)[38]and *Aegilops tauschi*(~2.6 Gb)[39]. TEs account for 74% and 77% for the A and D diploid genomesassemblies, respectively (Table S1).We also found a proportion of complete TEs of 12% and 11%, respectively, *i.e.* two times lower than on chromosome 3B (22%).

The annotation revealed that class I and II TEs represent 67% and 18% of the 3B sequence, respectively woth a vast majority of the chromosome is corresponding to LTR-RTs (529 Mb, i.e. 66% of the chromosome) (Table S1). Three superfamilies (*gypsy: 47%*, *CACTA:*16%, *copia:* 16%) account for more than 79% of the total TE fraction(Figure 1A). This proportion of *CACTAs* is much higher than in the other sequenced grasses: 3.2% in the maize[4], 4.7% in sorghum[12], 3.4% in rice[11], and 2.2% in *Brachypodiumdistachyon*[44]. In the draft genome sequences of *T. urartu,* and *Ae. tauschii*, we also found a higher proportion of *CACTAs*(15.6% and 12.3%respectively) (Table S1), suggesting that most of the *CACTA*amplification occurred before the divergence of the A, B and D genomes. Other DNA transposons are less abundant in terms of proportion but somecorrespond to small size elements found in very high copy numbers. For example, 17,479 miniature inverted-repeat transposable elements (MITEs; clustered into 95 families) of, on average,142bp, mostly from the *Mariner* superfamily, were detected along the 3B chromosome sequence.

In total, 485 families were detected along the 3B chromosome with only 6 families representing 50% of the TE fraction, and 143 representing 99% of the TE fraction (Figure 1B). The most abundant element was the LTR-RT RLG_fam1 (Fatima)family with 7036 complete and 8003 fragmented copies that account on their own for 12% of the TE fraction. Fifteen percent (74) of the families are represented by a single copy element on the entire chromosome 3B sequence while 2% (7) have amplified into more than 1000 copies. In contras to what was observed in maize, the majority of the TE families are not single copy member families but have rather amplified at a medium copy number (between 11 and 100) within this chromosome (Figure 1C). When calculating the distance between each TE and its neighboring genes we observed that low-copy number families are significantly closer to genes than highly repeated families (one-way anova, pvalues<$4.5e^{-11}$) (Figure 1D). Indeed, families with less than 100 copies were found to be significantly closer to genes than families having more than 100 copies (Bonnferroni/Dunn test, pvalues<$4e^{-4}$).

Overall, the TE distribution along the chromosome is strongly correlated with that of the *gypsy* elements (R=0.97) and is negatively correlated with the recombination rate (R=-0.82, pvalues<$10e^{-10}$) (Table S2). The *CACTA* superfamily exhibits the exact opposite pattern, with a significant increase of 23% in the distal regions (18-19%) compared to the proximal region (15%; Figure 2). Similar to *CACTAs*, but at a lower level, class II transposons *Harbingers*, *hATs*, *Mariners*, *Mutators*, and *Helitrons*, are twice as abundant in the distal regions compared to the proximal region and their distribution is strongly correlated with both the gene density (R=0.63) and recombination rate (R=0.71).

The distribution of TE diversity,*i.e.* the number of different families per10 Mb, along the chromosome was also investigated. It revealed a higher diversity in the distal regions (210 different TE families per 10 Mb on average, max: 244 families within 10 Mb) than in the central part of the chromosome (134 different families per 10 Mb; Figure S1). The diversity of DNA transposons increases by a 2.6 fold factor in the distal region compared to the centromere whereas the diversity in LTR-RTs is homogeneous along the chromosome (106+/-7 families per 10 Mb) (Figure S1). It is worth noting that the patterns of distribution of TE diversity and TE density are opposite. Thus, the increase in the amount of TEs in the centromeric regions is due to the accumulation of several copies from the same families.

Asegmentation analysis (see Material and Methods) of the TE density variations along the chromosome showed the presence of five distinct regions. The two chromosome ends, representing 18% of the chromosome (63 Mb and 73 Mb on the short and long arms, respectively),exhibit the lowest TE content (71%).The 122 Mb (16% of the chromosome region) encompassing the centromere[40]has the highest TE content of 93%. Finally, the two core parts of the chromosome arms (66% of the chromosome; 200 Mb and 316 Mb for the short and long arms, respectively) exhibit an average TE content of 88%. The borders of the two distal TE-poor regions correspond almost exactly to the regions (R1 and R3)defined by the recombination pattern in [40].

### *Uneven distribution and impact of TEs on the evolution of the chromosome structure*

To investigate the evolutionary dynamics of LTR-RTs in chromosome 3B, we estimated the insertion dates of 21,619 intact copies. We analyzed the distribution of insertion dates of individual copies for each of the 43 families with at least 20 copies in order to retrieve a family-specific burst date and period of activity (Figure 3A). It revealed that 93% of the TE bursts occurred between 1 and 3 MYA (Figure 3B), confirming at the whole chromosome scale, that TE amplification has been generally slowed down for the last 1 MY[31]. The active transposition periods (Figure 3B) lasted from 1 to 3 MY and correspond to a succession of bursts every 40,000 years on average. This indicates that the wheat genome has been shaped by successive waves of TE activation quickly followed by silencing. Indeed, 69% of the families were estimated to have been active over periods ranging from 1.5 to 2.5 MY. Only a few families were active for a longer period of time, suggesting they have escaped silencing. Not surprisingly, the intensity of the burst was negatively correlated with the amplification period (R=-0.4, pvalues=0.007), suggesting that the higher the level of activity the faster the silencing was established.

In order to investigate the type of evolutionary forces that have shaped the chromosome 3B structure, we studied the potential differences of chromosomal distribution of LTR-RT depending on

their insertion date. Four different categories were defined based on the TE insertion at <1 MYA, 1-2 MYA, 2-3 MYA, and >3 MYA (Figure 4). The analysis revealed that the gradual increase of the LTR-RT proportion from the telomeres to the centromere is in fact explained by an overrepresentation of ancient elements (>3 MYA). Indeed, no significant variation of the young (<1 Mya) LTR-RTs density was observed along the chromosome, whereas enrichments of 1.3, 2.3 and 2.8 fold were found in the proximal compared to the distal regions for classes 1-2 MY, 2-3 MY, and >3 MY, respectively. These results suggest that new TE insertions occur at a similar rate along the chromosome and, consequently, that the gradient observed is due to a higher rate of elimination in the recombinogenic distal regions.

One of the prevalent mechanisms for LTR-RT elimination is the formation of solo-LTR through ectopic homologous recombination between pairs of LTRs [16]. We detected 3,998 solo-LTRs with target site duplication (TSDs) on the chromosome 3B pseudomolecule, revealing a ratio of solo-LTRs to intact elements of 0.13:1, similar to what was observed previously in both wheat (0.14:1, [31]) and maize (0.14:1 [42]) but much lower than in rice (1.39:1; [42]). The second mechanism involved in TE turnover is illegitimate recombination that generates truncated TEs. A ratio of truncated to intact elements of 3.5:1 was found on the pseudomolecule, which is higher than the 0.5:1 ratio estimated previously for chromosome 3B [31]. This is mostly because of a higher number of gaps found in the assembly of the pseudomolecule *versus* a few BAC contigs whose sequence was completely finished that lead to predict intact elements as truncated. A higher ratio of truncated *versus* intact elements (27:1) was estimated in maize due to an even higher number of gaps, especially in LTR-RTs, in a sequence that was mostly aimed at completing the gene space [45]. Nevertheless, given that the assembly quality is homogeneous along chromosome3B, local variations of the truncated to intact ratio is investigated. It revealed a pattern very similar to the ratio of solo/intact elements (Figure S2) with a significant increase in the telomeric regions compared to the proximal region (one-way anova, pvalue<0.001), again suggesting a faster TE turnover in the recombinogenic regions. Hence, the proportions of both solo-LTRs and truncated LTR-RTs is correlated with the recombination rate (R=0.41 and R=0.56, respectively; pvalue<0.001).

Preferential insertion of large TEs could also play a role in the observed uneven distribution, since chromodomain-containing integrase of LTR-RTs could specifically target heterochromatin [13]. Among the 68 *gypsy* families identified on the 3B pseudomolecule, 7 encode a chromodomain-containing integrase and their distribution is biased with an increase in the proximal region(Figure S3). The abundance of *gypsy*harboring a chromodomain increases by 80% in the proximal region while *gypsy* without chromodomains increase by 20%, suggesting that chromatin affinity during transposition may also have contributed to this pattern.

### *Impact of CACTA transposons on the evolution of the gene content through gene duplication*

Chromosome 3B carries at least 2,065 genes/pseudogenes that are nonsyntenic with the related model grass genomes of *B. distachyon*, rice, and sorghum [40]. These genes originated from recent interchromosomal duplications (after the divergence with *Brachypodium*, 39 MYA [44]) and have preferentially accumulated in the distal regions of the chromosome. DNA transposon-mediated gene capture, which has already been suggested in wheat [31], was investigated to estimate the relationship between the increases of both *CACTAs* and duplicated genes in the distal regions. In order to identify *CACTA* families significantly associated with the nonsyntenic genes, a hierarchical clustering was applied on distributions of *CACTA* families (see Material and Methods; Figure 5A). We identified two groups containing 22 and 6 families (out of 30 families) with opposite patterns. The first group of 22 families was found to be overrepresented in the proximal region (e.g. *Jorge* family), while the second one containing 6 families was more abundant in the distal regions (e.g. *Caspar* family) (Figure 5B). Investigating potential site-specific insertion (within 50 bp around the elements) did not show any preferential insertion site for any of these groups. In addition, the two groups shared similar transposase domains. Distribution of subtelomeric-prone *CACTAs* was highly correlated with the distribution of nonsyntenic genes (R=0.8). Then, we compared their frequency in the close vicinity (+/-20 kb) of nonsyntenic *versus* syntenic genes (Figure 5C). It revealed that subtelomeric-prone *CACTAs* was found to be 2x higher in the vicinity of nonsyntenic *versus* syntenic genes. By contrast, no difference was observed for the centromeric-prone *CACTAs*. For example, the DTC_fam5 (*Vincent*)exhibited a 6-fold increase in the promoter (5 kb upstream) region of nonsyntenic genes compared to syntenic genes.

We detected 140 *CACTA*-mediated capture events involving 145 genes and 11 *CACTA* families (see Material and Methods). The DTC_fam5 (*Vincent*) family accounted for 74% (104) of the cases. Captured genes were smaller (average coding sequence (CDS) size 488 bp vs 1090 bp) and with fewer exons (3.05 vs 4.11) than the average genes on chromosome 3B. Forty-six percent of them exhibited a structure likely to be functional while the others were classified as pseudogenes. Using expression data (RNA-Seq performed on 5 organs at 3 developmental stages each; [46]), we showed that 26% (38) of the *CACTA*-captured genes were transcribed. The 145 gene copies captured represent 121 different families of genes. Interestingly, we found two cases where the same gene was captured twice by two different *CACTAs*. A putative 3'-5' exonuclease encoding gene is repeated in three copies of the DTC_fam5 (*Vincent*) (TRAES3BF090200030CFD_t1, TRAES3BF060400250CFD_t1, TRAES3BF060000130CFD_t1) and in seven copies of DTC_fam6 family (*TAT*) (TRAES3BF032300010CFD_t1, TRAES3BF032400050CFD_t1,

TRAES3BF067500010CFD_t1, TRAES3BF077700050CFD_t1, TRAES3BF082000030CFD_t1, TRAES3BF168400140CFD_t1, TRAES3BF182300010CFD_t1). The high level of conservation between the captured segments suggests that these two families have recently exchanged DNA.

We were able to detect 17 (12%) captured genes that potentially originate from exon shuffling involving 36 parental genes (see Material and Methods). Among them, 8 genes showed expression in at least in one of the conditions analyzed. Then, we estimated the type of selection pressure applied to those genes likely functional by estimating the dN/dS ratio ($\omega$) through sequence alignment with their closest homolog in *B. distachyon*. It revealed that most of the captured genes are under purifying selection, with a dN/dS ratio ranging from 0.2 to 0.6 (Figure 6). In contrast to the $\omega$ distribution observed for syntenic and nonsyntenic genes, we observed significantly larger dispersion for the captured genes (Fisher test, pvalues<$10^{-5}$), which indicates a more relaxed selection pressure on captured genes. In total, 19 captured genes (13%) were found to be expressed with a dN/dS ratio lower than 0.4.

Based on our finding on the 3B chromosome, we estimated that to ~2500 *CACTA*-captured gene/pseudogenes would be found at the whole genome scale, a range similar to the 1,194 genes captured by *Helitron* in maize [21] and 2,809 by *Pack-MULE* in rice [23].

## Discussion

### *A fine-tuned strategy for delivering accurate TE models in a highly complex genome*

Annotating TEs automatically and precisely in complex genomes is a challenge.However, a high-quality TE annotation is necessary, not only for performing evolutionary analyses and better understanding the impact of TEs on genome organization and expression, but also to prevent mis-prediction of genes with cellular function [29]. Here, we developed a strategy to overcome some of the problems due to the over-fragmentation of the predictions usually observed during TE modeling in large genomes. We used the knowledge accumulated during the past decades [31, 34, 36, 47] to fine-tune our algorithm and automatically reconstruct nested clusters that can still be identified. With CLARI-TE, we were able to predict a set of 56,488 intact TE on chromosome 3B. Half of the large intact TEs (>5 kb) were initially predicted as several truncated fragments through similarity search, revealing that for the time being curation is absolutely required to precisely delineate TEs in complex genomes. Moreover, among the 21,165 insertions of LTR-RTs that were used to study the dynamics of insertion/deletion along chromosome 3B, 30% were nested and, thus, would have been missed without automated joining by CLARI-TE. The use of our program prevented biasing the data towards recent TEs insertionsas usually observed in signature-based approaches [10, 42]. Finally, with a curated TE library, this approach can be applied to any complex genomes (>50% TEs) with nested TEs.

The proportion of complete elements was lower than estimated previously in our pilot study on selected BACs from chromosome 3B[31]: they represent 22% and 6% for LTR-RTs and CACTAs, respectively, compared to 59% and 47% in theprevious manually curated annotation. Similar observations have already been reported in maize where the ratio of truncated to intact elements was estimated to 0.5:1 [48] while it was 27:1 in the annotation of the reference genome sequence [45]. In both cases, the proportion of gaps, which is a reflection of the quality of the assembly, is the main limiting factor. Indeed, there are still 40,459 gaps on the chromosome 3B pseudomolecule, of which 57% were includedin a TE by CLARI-TE, while the others prevented from recovering the surrounding TE structure. About 85% of the TEs surrounding a gap were annotated as truncated. In addition, *CACTAs* are the largest wheat TEs (up to 30 kb), are highly variable, and contain tandem repeated motifs, features that make them the most challenging superfamily to identify automatically.

Estimate of the TE composition or organization of complex genomes are highly dependent on the reference TE library used as a basis for similarity search. In this study, similarity search-based annotation led us to assign 85% of the 3B sequence to TEs. Then, a *de novo* repeat identification allowed us to classify an additional 3% of the sequence to newly discovered repeats. Such a low

value revealed that the ClariTeRep library is an almost exhaustive representation of the TE diversity present on the 3B chromosome, suggesting that downstream analyses were not biased by a lack of knowledge regarding the TE composition of the chromosome. Although the three homoeologous wheat subgenomes share similar size and, thus, similar proportion of TEs, this proportion appeared substantially higher on chromosome 3B than the 66% and 67% estimates obtained for the diploid genomes of *Triticum urartu* (AA; [38]) and *Aegilops tauschii* (DD; [39]). Such a difference highlights the strong impact of the methodology used to annotate TEs, rather than a biological significance. This is supported by the results obtained after applying our TE modeling approach to these two draft genome sequences and the finding that TEs represent 74% and 77%, respectively, *i.e.*proportionsthat are closer to what we have found on 3B. The main difference is related to the lack of knowledge regarding *CACTAs* in the reference TE library. Here, we predicted 2 to 3 times more *CACTAs* than previously suggested (5.44 % for*T. urartu*, and 6.01 % for *Ae. tauschii*), confirming the impact of the reference TE library on the biological interpretations.Finally, differences of TE proportion found between the diploid genome and the chromosome 3Bis probably due to lack of sequence in the A and D draft genome sequence.With an estimated size of 5.5 Gb and 5 Gb for *T. urartu*[38]and *Ae. tauschii*[39, 49]respectively, TE should represent 80 to 82% of the genome, considering that low copy DNA represent 1 GB of the genome [40].

### *TE organization and dynamics*

The 774 Mb 3B pseudomolecule represents the largest chromosome sequence ever assembled in one single piece. The uneven distribution of recombination rate, gene density, gene expression pattern, and TEs along thechromosome highlighted a striking partitioning with five distinct regions: a centromeric/pericentromeric region with the highest TE density and in which recombination is suppressed; two subtelomeric regions with the lowest TE density and where recombination mainly occurs; two internal parts of chromosomal arms with intermediate features[40]. Behind the static view, the detailed study of TEs provided a dynamic view and novel insights into the evolutionary forces that have shaped this partitioning. First, we observed that LTR-RTs are the main contributors to the uneven distribution of TEs along the chromosome, a common feature of complex genomes like the maize genome[45]. In contrast to other grass genomes where TEs have mainly transposed recently, we found a major amplification period 1.5 MYA that has been followed by a period of silencing from 1.0 MYA until now, confirming that the two hybridization events that led to hexaploid wheat did not trigger massive activation of transposition. Previous study in wheat have observed a TE transposition burst immediately after allopolyploidization [36, 50], but as suggest by Parisod et al., it probably be a phenomenon restricted to specific TE families, and to mostly young active TE populations[51]. In addition, the availability of 21,165 complete LTR-RTs allowed us to study the TE

activity at a scale never reached so far, revealing that each family had its own period of activity before being silenced after 1 MY on average. Thus, overall, the B genome has been shaped by a succession of transposition waves from different families. Some of them escaping silencing rather than by massive reactivation of all TEs simultaneously. Similar waves of amplification have already been describe, such as in soybean [42], and are obviously observed in genome where LTR-RT have not been quickly eliminated from their host genome.

The distribution of the TE insertion time along the chromosome showed an uneven pattern that is mainly due to the differential location of old LTR-RTs (>3 MYA). In contrast, recently inserted elements (<1 MYA) exhibited a much more even distribution suggesting that transposition occurred at similar rate along the chromosome and that the decrease of TE density towards the telomeres rather reflects a rapid elimination in the recombinogenic distal regions. Similar conclusions were suggested for the sorghum [12]and maize [45] genomes, while for more compact genomes such as the one of*Arabidopsis*, the enrichment of the centromeric regions in LTR-RTs has been explained by a selection against the insertion of disruptive TEs in gene-rich regions [52]. Additional pieces of evidence for the more rapid elimination of TEsat the chromosome ends were provided by the overrepresentation of solo-LTRs and truncated LTR-RTs.This suggests that unequal homologous recombination (generating solo-LTRs) and illegitimate recombination (generating truncated TEs) are more frequent in the distal recombinogenic regions, as observed in rice [18]. In contrast to LTR-RTs, the density of class II DNA transposons increased in the distal regions and is positively correlated with both the gene density and recombination rate. Such correlation was also observed in sorghum [12] or maize [4] but not in rice [18]. Non-autonomous DNA transposons are well known to be associated with genes [53, 54] and their role in the regulation of gene expression has been suggested [8, 55]. In wheat, with the exception of *CACTAs*, DNA transposons are shorter than class I elements and their insertion into gene-rich regions might be counter-selected at a lower frequency. In addition, the faster TE turnover in the distal regions suggested above could also explain the increase of the DNA transposon density in these regions as a simple consequence of the compaction of the intergenic space that is mostly shaped by LTR-RTs. This differential deletion rate may also explain the increased TE diversity. Such pattern was described in maize where the centromere was perceived as an environment settled only by “individuals” the most adapted to proliferate, creating a diversity-poor ecosystem [45]. Finally, preferential TE insertion appeared as a potential driving force towards the partitioning of chromosome 3B with seven *gypsy* families having a chromodomain-containing integrase [13]that significantly concentrate in the proximal region.

***Impact of CACTA transposons on gene duplication in wheat***

DNA transposons represent 18% of the 3B chromosome sequence, which is the highest proportion observed among the sequenced grass genomes so far. *CACTAs* are the main contributors representing 16% while they only account for 2.2% to 5.9% in sorghum, rice, *Brachypodium,* maize and barley[4, 11, 44, 56].This supports the hypothesis that CACTAshave been amplified specifically in the wheat lineage. Generally, genome size is mainly correlated with class I elements, e.g. *LINE* in Human [57] and LTR-RTs in plants [58, 59], because of their copy-and-paste transposition mechanism allowing an increase in number in short time period [60, 61]. Thus, among plants, the wheat genome appears as a rare example of a massive amplification of cut-and-paste transposons.

Interestingly, 22 *CACTA* families were found to be preferentially associated with nonsyntenic genes i.e., genes that have been relocated via recent duplication events to a new chromosomal location. Previous studies in soybean [62], sorghum [12], or in *Ipomoea tricolor*[63] have shown that *CACTAs*can capture genes and gene fragments. Here, we identified 145 *CACTA*-captured genes on chromosome 3B (2% of the gene content). Although most of them were gene fragments, as already observed at a smaller scale [64], 13% were both transcribed and under purifying selection, suggesting they are functional, and 11%likely originating from exon shuffling. Beyond gene capture mechanisms, *CACTAs* were proven to mediate gene duplication in wheat through double strand break repair created at the time of insertion [64]. Preferential association of *CACTAs* with nonsyntenic genes might reflect a higher rate of gene duplication, due to the high frequency of *CACTA* insertion. This study highlights the importance of *CACTAs* on gene duplication and the creation of new genes that may be associated with the adaptation of wheat to various environments[40]. Similar examples involving class II transposons have been described with *Mutators* in rice [65][23] and *Helitrons* in maize [21, 22]. Therefore, it seems that superfamilies involved in gene capture have a tendency to proliferate and to be evolutionary successful.

## Conclusion

In this study, we annotated TEs and their nested pattern in one of the most complex genomes. Our automated proceduresignificantly improved the accuracy of TE predictions compared to a classical similarity-search approach. Such a high quality annotation enabled us to determine analyze the pattern of TE insertion, diversity and insertion time and revealed that the partitioning of the chromosome is mainly governed by higher deletion rates that are faster in recombining regions. We unraveled an unexpected abundance of CACTAs, and found a significant association with recently duplicated genes, suggestinga major impact of these elements on genome plasticity via the creation of genes. Such a mechanism may have provided wheat with an advantageous capability of adapting to wide range of environments.

## Material and Methods

### *Establishment of a classified library of TE sequences of* Triticeae

A library of TE sequences dedicated to similarity-search based annotation of TEs in the wheat genome was built as described in Choulet et al. 2014[40]. Briefly, we retrieved 3159 known full-length TE sequences i.e., elements having terminal repeats (terminal inverted repeat (TIRs) or LTRs) and/or features typical from complete *LINEs*/*SINEs*. We built 16 groups corresponding to each superfamily, and small non-autonomous MITEs were grouped separately from their autonomous counterparts to avoid computing multiple alignments with sequences of very different sizes. For each of the resulting 16 groups, an all-by-all BLAST [66] comparison (without filtering out low complexity sequences) was performed. BLAST output was analyzed with MCL (option -I 1.2) [67] in order to build clusters of sequences sharing similarity. The –I option, controlling the cluster granularity, was set to 1.2 for "*very coarse grained clustering*" meaning that large clusters were built at that stage. These clusters were used to define the family level. Families of 3 or more members were considered for computing a multiple alignment using MAFFT (default parameter) [68].A manual curation step was then applied. We used Jalview[69] as visualization tool for the manual curation of the multiple alignments and their corresponding neighbor joining tree. Sequences introducing mistakes in the multiple alignments (due to inversions, deletions, or insertions) were identified and discarded so that all alignments were corrected. In addition, since MCL grouped sequences within large clusters, we identified the clearly separated monophyletic groups (according to the neighbor joining tree) among each individual family and, therefore, defined variants within the family. For instance, family RLG_famc8 is comprised of 3 variants called (RLG_famc8.1, RLG_famc8.2, RLG_famc8.3). The library was called ClariTeRep and is available upon request.

### *Estimating the accuracy of TE prediction*

We used a scaffold of 904 kb from the wheat chromosome 3B that do not correspond to a previously known region of the genome as a test sequence to estimate the accuracy of the TE modeling. This reference scaffold carries 196 TEs covering 91% of the sequence, with 47 nested insertions. To automatize the comparison of the automated *versus* manually curated TE predictions and the calculation of sensitivity (Sn) and specificity (Sp) values, we developed compareAnnotTE.pl. Sensitivity and specificity were both estimated at three different levels: nucleotide, feature, and nested feature. At the nucleotide level, each nucleotide was considered to calculate Sn and Sp. At the feature level, only TE borders (all segments for TE split into several pieces by nested insertions) were considered.At the nested feature level,the program considers borders of nested TEs to

estimate the accuracy of the reconstructionof nested clusters. At the feature and nested feature levels, a predicted feature was considered as true positive if its borders correspond tothe manually curated TE positions in a range of 10 bp.

### *Similarity-search and automated curation using CLARI-TE*

We applied our procedure to the 2,808 scaffolds assembled for the 3B chromosome [HG670306 and CBUC010000001-CBUC010001450], the *Triticum urartu*[38], the *Aegilops tauschii* genomes[39], and 513 Mb of barely BAC clones [56] (ftp://ftpmips.helmholtz-muenchen.de/plants/barley/public_data/sequences/). Each sequencewasinvestigated for TE content using RepeatMasker (cross_match engine with default parameters; http://www.repeatmasker.org/) with ClariTeRep. We developed CLARI-TE to correct the raw similarity search results. It performs the three following steps. 1. Resolution of overlapping predictions. To solve the overlap between two predictions, priority was given to keep the prediction that covers an extremity of a TE. If none or both of the predictions cover a TE extremity, priority was given to keep the longest prediction and recalculate positions of the other one. 2. Merging predictions. Fragmentation of the TE models is dueto the presence of gaps in the scaffolds andto the fact that a newly identified TE copy may diverge from the reference element so that one element is not predicted as a single piece by is rather split into several pieces matching different parts of elements from the same family. In that case, all neighbor pieces related to the same family were merged into a single feature if the collinearity of the matching segments was respected, except for LTR matching segments. LTR positions of reference TEs were annotated in our library and this information was considered during the merging process. 3. Reconstruction of nested TEs. We developed a procedure to join separated features that are part of the same TE and have been split by nested insertions. Joining was allowed when 2 segments matching the same family (with respect of the collinearity between the prediction and the reference TE) are separated by at maximum 10 predicted TEs. The final stage of the annotation is the assignment of intact full-length *versus* fragmented TEs. Intact full-length TEs are predictions covering at least 90% of the reference complete TE in the library and for which both extremities were identified (in a range of 50 nucleotides). Moreover, PFAM domains were search for every complete element to detect chromo-domains and transposase-likedomains.

### *Estimation of LTR-RT insertion date*

We used the program TRsearch from REPET [43]to find position of both 5' and 3' LTRs from a complete element. We discarded predicted LTRs that did not correspond to the extremity of an element (in a range of 50 bp). Pairs of LTRs were aligned using MUSCLE [70] and insertion dates of LTR-RTs were estimated considering a mutation rate of $1.3*10^{-8}$ substitutions/site/year [71].

Finally, using R, the distribution of insertion dates was plotted for each family accounting 20 or more copies with an estimated date. For each distribution, a burst peak date was determined and a period of activity was calculated by considering the shortest period of time containing more than 80% of the dated insertions.

### *Distribution of TEs along the chromosome*

Distribution of TEs along the chromosome 3B sequence was computed by calculating the proportion (in size) and number of TEs in a sliding window of 10 Mb with a step of 1 Mb. The distribution of the TE diversity along the chromosome was computed using the same window by calculating the number of families per window. To prevent from drastic variations due to mis-predictions, we considered the number of families representing 99% of the TE fraction per window (N99).

### *Prediction of solo-LTRs*

Based on of the 30,406 intact LTR-RTs predicted on chromosome 3B, we built a library of LTR sequences by extracting 18,928 LTRs flanked by canonical 5'-TG and 3'-CA dinucleotides. This library was used for an additional round of similarity search using RepeatMasker on the full chromosome 3B. To distinguish solo LTRs from truncated LTR-RTs, we searched specifically for the presence of a 5 bp TSD (one nucleotide variation tolerated) flanking the matching region.

### *Hierarchical clustering of distributions*

In order to identify *CACTA* families with similar distributions along the chromosome, we performed a hierarchical clustering of the distributions calculated in a sliding window of 10 Mb (step: 1 Mb) by using the R package ‘pvclust’ [72]. We considered only families representing at least 0.01% in at least one 10 Mb window. The pearson correlations were calculated between each pairs of distribution and a clustering was applied by the agglomerative method "average" (N=10000 bootstrap resampling).

### *TE abundance in the vicinity of genes*

The CDS positions of the annotated chromosome 3B protein-coding genes was used to estimate the relative abundance of TE families in the 20 kb upstream and downstream regions. The average proportion of a given TE family was calculted for each nucleotide position in the CDS sourrouding sequence by considering the orientation of the genes.

### *Detection of TE-captured gene and exon shuffling events*

In order to detect TE-captured genes, we isolated genes flanked by 2 elements belonging to the same family, as a trace of potential capture event. A total of 558 potential gene capture events were detected, 235 (42%) involving a *CACTA*, 104 (19%) involving a *gypsy*, and 219 (39%) involving other superfamilies. We manually checked the presence of target site duplications as an evidence of gene capture.

To decipher potential exon shuffling events, we searched for the presence of chimeric genes. A similarity search using BLASTP against *B. distachyon* proteome was performed for each captured gene product. Wheat proteins aligning over 70% of their length to a *B. distachyon* protein were filtered out. For the others, we develop a parsing procedure to solve the overlapping regions of similarity between a *B. distachyon* protein and our query sequence and to detect chimeric proteins i.e. proteins showing non-overlapping segments of similarity with different *B. distachyon* proteins.

### *Detection of duplicated copies of TE-captured genes at the whole genome scale*

All copies of captured genes initially identified on chromosome 3B were searched among the chromosome survey sequences of the hexaploid genome [41]. BLASTN was used to search for similarity with the captured CDSs and with their 5' and 3' junctions between the captured CDSs and their associated *CACTA*. These junctions were defined as from 100 bp inside the CDS to 100 bp inside the *CACTA* conserved sequence. Survey sequence contigs showing similarity with both the CDS and at least one of these junctions were considered as additional captured copies. Exonerate was then used to precisely annotate the structure of the CDS in the chromosome survey sequence contigs [73].

## List of Abbreviations used

TE: Transposable Element, LTR-RT: Long Terminal Repeat Retrotransposon, MITE: Miniature Inverted-Repeat Transposable Element, TSD: Target Site Duplication, CDS: Coding Sequence, TIR: Terminal Inverted Repeat

## Competing Interests

The authors declare that they have no competing interests.

## Authors Contribution

JDperformed programming, conducted data analyses and wrote the paper. NG undertakes the nonsyntenic gene detection. LP collected and analyzed RNAseq expression data and carried out the cluster analysis.ST made the gene annotation and genome assembly. VJ participated to the TE annotation process. EP and HQ participated in the design of the study and critical revision of the manuscript. VB, SM, AA, PW were involved in producing the sequence of chromosome 3B, CF participated in the design of the project, data interpretation, coordination of the project, and critical revision of the manuscript. FCdesigned the research, supervised this work, participated to data analyses and finalized the paper.

## Acknowledgements

This project was supported by grants from the French National Research Agency (ANR-09-GENM-025 3BSEQ), a grant of France Agrimer, and a grant (project DL-BLE) from the INRA BAP division.JD is funded by a grant from the French Ministry of Research.NG is funded by a grant of the European Commission research training program Marie-Curie Actions (FP7-MC-IIF-NoncollinearGenes). LP is funded by a grant from the Region Auvergne.

## References

1. McClintock B: **Mutable loci in maize**, vol. 47: Year Book of the Carnegie Institution of Washington; 1948.
2. Feschotte C, Jiang N, Wessler SR: **Plant transposable elements: where genetics meets genomics**. *Nature Reviews Genetics* 2002, **3**(5):329-341.
3. Arabidopsis GI: **Analysis of the genome sequence of the flowering plant Arabidopsis thaliana**. *Nature* 2000, **408**(6814):796.
4. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA: **The B73 maize genome: complexity, diversity, and dynamics**. *science* 2009, **326**(5956):1112-1115.
5. Devos KM: **Grass genome organization and evolution**. *Current opinion in plant biology* 2010, **13**(2):139-145.

6. Bennetzen JL, Ma J, Devos KM: **Mechanisms of recent genome size variation in flowering plants**. *Annals of botany* 2005, **95**(1):127-132.
7. Tenaillon MI, Hollister JD, Gaut BS: **A triptych of the evolution of plant transposable elements**. *Trends in plant science* 2010, **15**(8):471-478.
8. Naito K, Zhang F, Tsukiyama T, Saito H, Hancock CN, Richardson AO, Okumoto Y, Tanisaka T, Wessler SR: **Unexpected consequences of a sudden and massive transposon amplification on rice gene expression**. *Nature* 2009, **461**(7267):1130-1134.
9. Li B, Choulet F, Heng Y, Hao W, Paux E, Liu Z, Yue W, Jin W, Feuillet C, Zhang X: **Wheat centromeric retrotransposons: the new ones take a major role in centromeric structure**. *The Plant Journal* 2013, **73**(6):952-965.
10. Vitte C, Bennetzen JL: **Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution**. *Proceedings of the National Academy of Sciences* 2006, **103**(47):17638-17643.
11. Matsumoto T, Wu J, Kanamori H, Katayose Y, Fujisawa M, Namiki N, Mizuno H, Yamamoto K, Antonio BA, Baba T: **The map-based sequence of the rice genome**. *Nature* 2005, **436**(7052):793-800.
12. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A: **The Sorghum bicolor genome and the diversification of grasses**. *Nature* 2009, **457**(7229):551-556.
13. Gao X, Hou Y, Ebina H, Levin HL, Voytas DF: **Chromodomains direct integration of retrotransposons to heterochromatin**. *Genome Research* 2008, **18**(3):359-369.
14. Marín I, Lloréns C: **Ty3/Gypsy retrotransposons: description of new Arabidopsis thaliana elements and evolutionary perspectives derived from comparative genomic data**. *Molecular biology and evolution* 2000, **17**(7):1040-1049.
15. Chatterjee AG, Leem YE, Kelly FD, Levin HL: **The chromodomain of Tf1 integrase promotes binding to cDNA and mediates target site selection**. *Journal of virology* 2009, **83**(6):2675-2685.
16. Devos KM, Brown JK, Bennetzen JL: **Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis**. *Genome Research* 2002, **12**(7):1075-1079.
17. Ma J, Devos KM, Bennetzen JL: **Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice**. *Genome Research* 2004, **14**(5):860-869.
18. Tian Z, Rizzon C, Du J, Zhu L, Bennetzen JL, Jackson SA, Gaut BS, Ma J: **Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons?***Genome Research* 2009, **19**(12):2221-2230.
19. Duret L, Marais G, Biémont C: **Transposons but not retrotransposons are located preferentially in regions of high recombination rate in Caenorhabditis elegans**. *Genetics* 2000, **156**(4):1661-1669.
20. Long M, Betrán E, Thornton K, Wang W: **The origin of new genes: glimpses from the young and old**. *Nature Reviews Genetics* 2003, **4**(11):865-875.
21. Yang L, Bennetzen JL: **Distribution, diversity, evolution, and survival of Helitrons in the maize genome**. *Proceedings of the National Academy of Sciences* 2009, **106**(47):19922-19927.
22. Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A: **Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize**. *Nature genetics* 2005, **37**(9):997-1002.
23. Hanada K, Vallejo V, Nobuta K, Slotkin RK, Lisch D, Meyers BC, Shiu S-H, Jiang N: **The functional role of pack-MULEs in rice inferred from purifying selection and expression profile**. *The Plant Cell Online* 2009, **21**(1):25-38.
24. Jiang N, Wessler SR: **Insertion preference of maize and rice miniature inverted repeat transposable elements as revealed by the analysis of nested elements**. *The Plant Cell Online* 2001, **13**(11):2553-2564.
25. Juretic N, Bureau TE, Bruskiewich RM: **Transposable element annotation of the rice genome**. *Bioinformatics* 2004, **20**(2):155-160.
26. Jiang N, Feschotte C, Zhang X, Wessler SR: **Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs)**. *Current opinion in plant biology* 2004, **7**(2):115-119.
27. Lu C, Chen J, Zhang Y, Hu Q, Su W, Kuang H: **Miniature Inverted–Repeat Transposable Elements (MITEs) Have Been Accumulated through Amplification Bursts and Play Important Roles in Gene Expression and Species Diversity in Oryza sativa**. *Molecular biology and evolution* 2012, **29**(3):1005-1017.

28. Chen J, Hu Q, Zhang Y, Lu C, Kuang H: **P-MITE: a database for plant miniature inverted-repeat transposable elements**. *Nucleic acids research* 2013:gkt1000.
29. Bennetzen JL, Coleman C, Liu R, Ma J, Ramakrishna W: **Consistent over-estimation of gene number in complex plant genomes**. *Current opinion in plant biology* 2004, **7**(6):732-736.
30. Lerat E: **Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs**. *Heredity* 2009, **104**(6):520-533.
31. Choulet F, Wicker T, Rustenholz C, Paux E, Salse J, Leroy P, Schlub S, Le Paslier M-C, Magdelenat G, Gonthier C: **Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces**. *The Plant Cell Online* 2010, **22**(6):1686-1701.
32. SanMiguel PJ, Ramakrishna W, Bennetzen JL, Busso CS, Dubcovsky J: **Transposable elements, genes and recombination in a 215-kb contig from wheat chromosome 5Am**. *Functional & integrative genomics* 2002, **2**(1-2):70-80.
33. Li W, Zhang P, Fellers JP, Friebe B, Gill BS: **Sequence composition, organization, and evolution of the core Triticeae genome**. *The Plant Journal* 2004, **40**(4):500-511.
34. Sabot F, Guyot R, Wicker T, Chantret N, Laubin B, Chalhoub B, Leroy P, Sourdille P, Bernard M: **Updating of transposable element annotations from large wheat genomic sequences reveals diverse activities and gene associations**. *Molecular Genetics and Genomics* 2005, **274**(2):119-130.
35. Paux E, Roger D, Badaeva E, Gay G, Bernard M, Sourdille P, Feuillet C: **Characterizing the composition and evolution of homoeologous genomes in hexaploid wheat through BAC□end sequencing on chromosome 3B**. *The Plant Journal* 2006, **48**(3):463-474.
36. Charles M, Belcram H, Just J, Huneau C, Viollet A, Couloux A, Segurens B, Carter M, Huteau V, Coriton O: **Dynamics and differential proliferation of transposable elements during the evolution of the B and A genomes of wheat**. *Genetics* 2008, **180**(2):1071-1086.
37. Brenchley R, Spannagl M, Pfeifer M, Barker GL, D'Amore R, Allen AM, McKenzie N, Kramer M, Kerhornou A, Bolser D: **Analysis of the bread wheat genome using whole-genome shotgun sequencing**. *Nature* 2012, **491**(7426):705-710.
38. Ling H-Q, Zhao S, Liu D, Wang J, Sun H, Zhang C, Fan H, Li D, Dong L, Tao Y: **Draft genome of the wheat A-genome progenitor Triticum urartu**. *Nature* 2013.
39. Jia J, Zhao S, Kong X, Li Y, Zhao G, He W, Appels R, Pfeifer M, Tao Y, Zhang X: **Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation**. *Nature* 2013.
40. Choulet F, Alberti A, Theil S, Glover N, Barbe V, Daron J, Pingault L, Sourdille P, Couloux A, Paux E *et al*: **Structural and functional partitioning of bread wheat chromosome 3B**. *science* 2014, **345**(6194):1249721.
41. Mayer KF, Rogers J, Doležel J, Pozniak C, Eversole K, Feuillet C, Gill B, Friebe B, Lukaszewski AJ, Sourdille P: **A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome**. *science* 2014, **345**(6194):1251788.
42. El Baidouri M, Panaud O: **Comparative Genomic Paleontology across Plant Kingdom Reveals the Dynamics of TE-Driven Genome Evolution**. *Genome biology and evolution* 2013, **5**(5):954-965.
43. Flutre T, Duprat E, Feuillet C, Quesneville H: **Considering transposable element diversification in de novo annotation approaches**. *PLoS One* 2011, **6**(1):e16526.
44. Vogel JP, Garvin DF, Mockler TC, Schmutz J, Rokhsar D, Bevan MW, Barry K, Lucas S, Harmon-Smith M, Lail K: **Genome sequencing and analysis of the model grass Brachypodium distachyon**. *Nature* 2010, **463**(7282):763-768.
45. Baucom RS, Estill JC, Chaparro C, Upshaw N, Jogi A, Deragon J-M, Westerman RP, SanMiguel PJ, Bennetzen JL: **Exceptional Diversity, Non-Random Distribution, and Rapid Evolution of Retroelements in the B73 Maize Genome**. *PLoS Genet* 2009, **5**(11):e1000732.
46. Rustenholz C, Choulet F, Laugier C, Šafář J, Šimková H, Doležel J, Magni F, Scalabrin S, Cattonaro F, Vautrin S: **A 3,000-loci transcription map of chromosome 3B unravels the structural and functional features of gene islands in hexaploid wheat**. *Plant physiology* 2011, **157**(4):1596-1608.
47. Wicker T, Guyot R, Yahiaoui N, Keller B: **CACTA transposons in Triticeae. A diverse family of high-copy repetitive elements**. *Plant physiology* 2003, **132**(1):52-63.
48. Liu R, Vitte C, Ma J, Mahama AA, Dhliwayo T, Lee M, Bennetzen JL: **A GeneTrek analysis of the maize genome**. *Proceedings of the National Academy of Sciences* 2007, **104**(28):11844-11849.
49. Rees H, Walters M: **Nuclear DNA and the evolution of wheat**. *Heredity* 1965, **20**(1):73-82.
50. Kashkush K, Feldman M, Levy AA: **Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat**. *Nature genetics* 2002, **33**(1):102-106.

51. Parisod C, Alix K, Just J, Petit M, Sarilar V, Mhiri C, Ainouche M, Chalhoub B, Grandbastien MA: **Impact of transposable elements on the organization and function of allopolyploid genomes**. *New Phytologist* 2010, **186**(1):37-45.
52. Wright SI, Agrawal N, Bureau TE: **Effects of recombination rate and gene density on transposable element distributions in Arabidopsis thaliana**. *Genome Research* 2003, **13**(8):1897-1903.
53. Wessler S: **Transposable elements and the evolution of gene expression**. In: *Symposia of the Society for Experimental Biology: 1997*; 1997: 115-122.
54. Han Y, Qin S, Wessler SR: **Comparison of class 2 transposable elements at superfamily resolution reveals conserved and distinct features in cereal grass genomes**. *BMC genomics* 2013, **14**(1):71.
55. Yang G, Lee Y-H, Jiang Y, Shi X, Kertbundit S, Hall TC: **A two-edged role for the transposable element Kiddo in the rice ubiquitin2 promoter**. *The Plant Cell Online* 2005, **17**(5):1559-1568.
56. Consortium IBGS: **A physical, genetic and functional sequence assembly of the barley genome**. *Nature* 2012, **491**(7426):711-716.
57. Batzer MA, Deininger PL: **Alu repeats and human genomic diversity**. *Nature Reviews Genetics* 2002, **3**(5):370-379.
58. Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF: **Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium**. *Genome Research* 2006, **16**(10):1252-1261.
59. Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, Collura K, Brar DS, Jackson S, Wing RA: **Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice**. *Genome Research* 2006, **16**(10):1262-1269.
60. Vitte C, Panaud O: **LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model**. *Cytogenetic and genome research* 2005, **110**(1-4):91-107.
61. Hawkins JS, Proulx SR, Rapp RA, Wendel JF: **Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants**. *Proceedings of the National Academy of Sciences* 2009, **106**(42):17811-17816.
62. Zabala G, Vodkin LO: **The wp mutation of Glycine max carries a gene-fragment-rich transposon of the CACTA superfamily**. *The Plant Cell Online* 2005, **17**(10):2619-2632.
63. Takahashi S, Inagaki Y, Satoh H, Hoshino A, Iida S: **Capture of a genomic HMG domain sequence by the En/Spm-related transposable element Tpn1 in the Japanese morning glory**. *Molecular and General Genetics MGG* 1999, **261**(3):447-451.
64. Wicker T, Buchmann JP, Keller B: **Patching gaps in plant genomes results in gene movement and erosion of colinearity**. *Genome Research* 2010, **20**(9):1229-1237.
65. Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR: **Pack-MULE transposable elements mediate gene evolution in plants**. *Nature* 2004, **431**(7008):569-573.
66. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.
67. Van Dongen S: **A cluster algorithm for graphs**. *Report-Information systems* 2000(10):1-40.
68. Katoh K, Kuma K-i, Toh H, Miyata T: **MAFFT version 5: improvement in accuracy of multiple sequence alignment**. *Nucleic acids research* 2005, **33**(2):511-518.
69. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ: **Jalview Version 2—a multiple sequence alignment editor and analysis workbench**. *Bioinformatics* 2009, **25**(9):1189-1191.
70. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucleic acids research* 2004, **32**(5):1792-1797.
71. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL: **The paleontology of intergene retrotransposons of maize**. *Nature genetics* 1998, **20**(1):43-45.
72. Suzuki R, Shimodaira H: **Pvclust: an R package for assessing the uncertainty in hierarchical clustering**. *Bioinformatics* 2006, **22**(12):1540-1542.
73. Slater GS, Birney E: **Automated generation of heuristics for biological sequence comparison**. *BMC bioinformatics* 2005, **6**(1):31.

## Tables

**Table 1: Comparison of the accuracy of TE modeling on a 1 Mb scaffold of wheat chromosome 3B.**The reference annotation was curated manually. Similarity search was performed using RepeatMasker (RM) and automated TE modeling based on the RM results was performed using either TEannot or CLARI-TE. The sensitivity (sn) and specificity (sp) were calculated at three different levels: nucleotide, feature and nested feature (see Materials and Methods).

| | | reference | RM | TEannot | CLARI-TE |
|---|---|---|---|---|---|
| # predictions | | 196 | 590 | 345 | 289 |
| coverage | | 91% | 90% | 90% | 91% |
| nucleotide | Sn | - | 93% | 95% | 96% |
| | Sp | - | 96% | 97% | 95% |
| predictions | Sn | - | 54% | 45% | 66% |
| | Sp | - | 26% | 27% | 52% |
| nested TE | Sn | - | NA | 17% | 41% |
| | Sp | - | NA | 14% | 58% |

## Figure legends

**Figure 1: TE content and copy number of the wheat chromosome 3B sequence.** (A) Pie graph of the relative composition of the main TE superfamilies. (B) Cumulative sum of the number of TEs families among the TE fraction. The N50 is 6, meaning that 6 TE families represent 50% of the TE fraction (in number of nucleotides). (C) Distribution of the number of copies per family (considering complete copies only). (D) Box plot of the distance (in kb) of TEs to the closest gene. The 5 categories represent TE families with different number of copies on the 3B chromosome.

**Figure 2: Distribution of the variations of TE density along the wheat chromosome 3B.** Distributions are represented for four superfamilies: *gypsy* (blue), *copia* (green), *CACTA* (red), and other DNA transposons (purple). The distributions were calculated in a sliding window of 10 Mb with a step of 1 Mb and the graphs represent the standard score i.e., variations around the mean in number of standard deviations. Recombinogenic distal and centromeric/pericentromeric regions [40] are labeled by grey and hatched areas, respectively.

**Figure 3: Insertion date and period of amplification of LTR-RTs on wheat chromosome 3B.** (A) Distribution of the insertion dates for 5,554 complete copies of the family RLG_fam1 (Fatima). The number of copies is highlighted in green (top), the peak of amplification is in red (middle), and the period of activity is in blue (bottom). (B) Distribution of the frequency of the copy number, insertion dates, and period of activity using 43 LTR-RT families with at least 20 copies.

**Figure 4: Relationships between insertion dates and chromosomal location of LTR-RTs.** In total, 21,619 LTR-RTs with an estimated insertion date have been grouped within four classes: 0-1 MYA (blue), 1-2 MYA (purple), 2-3 MYA (green), >3 MYA (red). (A) Distribution of the number of LTR-RTs for each of these 4 categories along the chromosome 3B. Distribution was calculated in a sliding window of 10 Mb with a step of 1 Mb. Recombinogenic distal and centromeric/pericentromeric regions [40] are labeled by grey and hatched areas, respectively. (B) Box plot of the number of TEs per 10 Mb sliding window carried by the recombinogenic distal regions (D), the internal parts of the chromosome arms (I), and the centromeric/pericentromeric regions (C).

**Figure 5: Relationships between CACTA families and nonsyntenic genes.** (A) Tree based on the clustering of the distribution patterns of 36 *CACTA* families found along the 3B chromosome. The branches represented in red and blue correspond to the families that are overrepresented and underrepresented, respectively, in the distal regions of the chromosome. (B) Distribution of the 28 *CACTA* families (top) and 6 *CACTA* families (bottom) showing opposite patterns along the 3B chromosome sequence. The gray curves represent the distribution of individual families and the top blue or red curves represent the cumulative sum of all families. (C) Abundance of *CACTAs* in the vicinity (+/-20 kb) of syntenic (left panel) and nonsyntenic (right panel) genes. 0 represent the position of the CDSs (start and stop codons) and the average abundance of *CACTAs* was calculated for each nucleotide in a -20/+20 kb window encompassing the genes. Red and blue curves represent the relative abundance of the *CACTAs* overrepresented (6 families) and underrepresented (28 families), respectively, in the distal regions.

**Figure 6: Selection pressure estimated by the dN/dS ratio for CACTA-captured genes.** Distribution of the frequency of the dN/dS ratio for 2964 syntenic genes(blue), 1179 nonsyntenic genes (red) and 127 *CACTA*-captured genes on chromosome 3B (green).

**Annexe 4** : Thomas, M., **Pingault, L**., Poulet, A., Duarte, J., Throude, M., Faure, S., … Tatout, C. (s. d.). Evolutionary history of Methyltransferase 1 genes in hexaploid wheat.

BMC Genomics

RESEARCH ARTICLE Open Access

# Evolutionary history of Methyltransferase 1 genes in hexaploid wheat

Mélanie Thomas[1,2], Lise Pingault[3], Axel Poulet[1], Jorge Duarte[2], Mickaël Throude[2], Sébastien Faure[2], Jean-Philippe Pichon[2], Etienne Paux[3], Aline Valeska Probst[1] and Christophe Tatout[1*]

## Abstract

**Background:** Plant and animal methyltransferases are key enzymes involved in DNA methylation at cytosine residues, required for gene expression control and genome stability. Taking advantage of the new sequence surveys of the wheat genome recently released by the International Wheat Genome Sequencing Consortium, we identified and characterized *MET1* genes in the hexaploid wheat *Triticum aestivum* (*TaMET1*).

**Results:** Nine *TaMET1* genes were identified and mapped on homoeologous chromosome groups 2A/2B/2D, 5A/5B/5D and 7A/7B/7D. Synteny analysis and evolution rates suggest that the genome organization of *TaMET1* genes results from a whole genome duplication shared within the grass family, and a second gene duplication, which occurred specifically in the *Triticeae* tribe prior to the speciation of diploid wheat. Higher expression levels were observed for *TaMET1* homoeologous group 2 genes compared to group 5 and 7, indicating that group 2 homoeologous genes are predominant at the transcriptional level, while group 5 evolved into pseudogenes. We show the connection between low expression levels, elevated evolution rates and unexpected enrichment in CG-dinucleotides (CG-rich isochores) at putative promoter regions of homoeologous group 5 and 7, but not of group 2 *TaMET1* genes. Bisulfite sequencing reveals that these CG-rich isochores are highly methylated in a CG context, which is the expected target of *TaMET1*.

**Conclusions:** We retraced the evolutionary history of *MET1* genes in wheat, explaining the predominance of group 2 homoeologous genes and suggest CG-DNA methylation as one of the mechanisms involved in wheat genome dynamics.

**Keywords:** DNA methylation, Evolution, Genome dynamics, CG-rich isochores

## Background

*Triticum aestivum* (hexaploid wheat or bread wheat) is one of the most important cultivated species in the world and it has been subject of intense research. Investigations of its genome structure led to the discovery of its highly dynamic nature during evolution. Using fossil records and phylogenetic studies, its evolution was traced from ancestral diploid species proposed to originate 50–77 million years ago (MYa) [1]. Indeed, bread wheat is a hexaploid species made of three homoeologous genomes called A, B and D which derived from different diploid species. These are proposed to be *Triticum urartu* (2n = 2× = 14 chromosomes, AA) and a diploid species related to *Aegilops speltoides* (2n = 2× = 14, BB) which gave rise some 0.5 ~ 0.6 MYa ago to *Triticum durum* (2n = 4× = 28 chromosomes, AABB). About 8,000 years ago, hybridization occurred between *Triticum durum* and *Aegilops tauschii* (2n = 2× = 14 chomosomes, DD) and yielded *Triticum aestivum* (2n = 6× = 42 chromosomes, AABBDD), the hexaploid wheat [2]. This means that every single gene is expected to be found in triplicate. The genome structure, organized in homoeologous genomes A, B and D, has to be maintained through cell division, a function which is ensured by the *Ph1* suppressor locus. The *Ph1* locus restricts homoeologous recombination and permits proper chromosome segregation in a hexaploid context through mitosis and meiosis [3].

* Correspondence: christophe.tatout@univ-bpclermont.fr
[1]UMR CNRS 6293 INSERM U 1103 Clermont Université, Genetics Reproduction and Development (GReD), 24 avenue des Landais, BP80026, 63171 Aubière Cedex, France
Full list of author information is available at the end of the article

Complementary approaches known as comparative genomics [4] at the genome-level (synteny) or the chromosome level (micro-synteny) were used to predict the genome structure of wheat in comparison to sequenced diploid species such as rice [5,6], sorghum [7], maize [8], brachypodium [9] and more recently barley [10]. Recent syntenic studies proposed that the ancestral genome of grass species was a diploid species organized in five chromosomes (2n = 2× = 10 chromosomes) [11]. From this initial chromosome organization, the ancestral diploid genome was duplicated through Whole-Genome Duplication (WGD) then fragmented giving rise to an intermediate ancestor with 2n = 2× = 24 chromosomes [11]. This genomic structure has been well conserved in rice (2n = 2× = 24 chromosomes) while it evolved to 2n = 2× = 14 through chromosome rearrangements in diploid wheat. Although WGD is expected to have had a large impact on wheat genome evolution it is not the only mode of genome rearrangement. Indeed, duplication of large chromosomal regions (segmental duplication), duplication at the gene level or tandem duplications have also occurred in the course of evolution [12]. Furthermore, it is now well established that wheat genome organization has been largely influenced by transposable element mobilization [13]. Most of the mechanisms described above increase genome size and lead to an elevated gene copy number. However, much less is known about reverse mechanisms, which reduce genome size to restore a diploid situation and reform single copy gene states. Indeed, early studies in *Saccharomyces cerevisiae* indicate that only 12% of the duplicate pairs remain after WGD suggesting that an extensive gene loss occurs after WGD [14]. In flowering plants, a fraction of single-copy genes were recently investigated and new hypotheses were suggested [15]: basically, after duplication, genes within one of the duplicated segments tend to be lost through small deletions while most genes are retained within the second segment, a mechanism known as fractionation bias [16]. Another difference occurring after duplication between two genomic segments is known as genome dominance during which one of the two segments shows higher expression levels than the other [16]. Data from maize and brassica further indicate that both gene fractionation, leading to extensive gene loss, and genome dominance are occurring simultaneously keeping the expression of the retained genes at elevated levels [16,17]. Hexaploid wheat does not show an overall genome-wide transcriptional dominance of A, B or D subgenomes although some homoeologous genes can adopt a specific expression pattern [18]. All these recent outcomes reveal important genome dynamics, which affect genome size or organization and alter gene expression. However, mechanisms implicated in these phenomena remain largely hypothetical, although epigenetic mechanisms have been suggested to provide means to induce asymmetric levels of expression between the two duplicated fragments prior to gene fractionation [16].

Although our knowledge about the hexaploid wheat genome structure is increasing, it remains challenging to decipher every step leading to its large genome size of about 16–17 Gb, which includes up to 80% of repeated sequences [13]. In polyploid genomes like cotton, rapeseed or wheat, several studies suggested the importance of epigenetic mechanisms in maintaining genome structure and chromatin stability as well as in regulating gene expression after hybridization and polyploidization [19,20]. One of these epigenetic mechanisms is DNA methylation, which takes place at the carbon-5 cytosine residues in CG, CHG and CHH (where H = A, T or C) contexts [21]. Loss of DNA methylation causes reactivation of silenced transposable elements [22] and the expression of certain genes, such as *FWA*, a gene involved in flowering [23,24]. DNA methylation is also known to affect crossover rate and meiotic recombination [25].

We wanted to reconstruct the evolutionary history of the hexaploid wheat species *Triticum aestivum* using the example of *MET1*, a gene encoding METHYLTRANSFERASE 1 (MET1), responsible for DNA methylation maintenance in the CG context. *MET1* is a gene of particular importance for genome maintenance in many organisms, which we hypothesize will be a crucial component of epigenetic mechanisms controlling transposable elements that in wheat make up to 80% of the genome. To date *MET1* gene function have been described in several plant species including Arabidopsis [26], maize [27], rice [28] and brassica [29] but not in wheat. We identified *MET1* genes in hexaploid wheat (*TaMET1*). Nine copies of *TaMET1* are organized in three paralogous groups at chromosome 2, 5 and 7 suggesting that the genomic regions including *MET1* genes were subjected to two duplication events prior to the emergence of hexaploid wheat. Considering *TaMET1* genomic regions specifically, we confirmed that the first gene duplication was part of a WGD common to all grass species and that the second duplication occurred through gene duplication specific to the *Triticeae* tribe. Expression profiles of the different *MET1* gene copies, estimation of their evolution rates, CG enrichment and methylation profiles highlight the predominance of group 2 homoeologous genes at the transcript level. Our results exemplify the high dynamics of genome evolution in the course of the evolutionary history of bread wheat and suggest the involvement of epigenetic mechanisms in these processes.

## Results

### Hexaploid wheat contains nine *TaMET1* loci

In order to determine the number and complete sequence of *TaMET1* genes, we chose a genomic strategy based on

a combination of sequence capture and *in silico* mining of available wheat genome sequences. In order to define probes on the sequence capture microarray, *TaMET1*-expressed tags (ESTs) were identified in wheat databases. Eight ESTs were retrieved from public and private libraries. The alignment of these ESTs with the rice and brachypodium *MET1* genes showed that these ESTs mapped to the 3′end of *TaMET1* genes. Two *TaMET1* ESTs as well as two brachypodium *MET1* genomic fragments were selected and used to design probes for sequence capture (see Methods). Two successive runs of sequence capture yielded 8,184 reads specific to *TaMET1* genes. Genomic fragments were then assembled *de novo* using gsAssembler in six large contigs corresponding to six putative *TaMET1* genes. However some reads remained impossible to assemble and could not be included within the six large contigs suggesting the possible existence of additional copies of *TaMET1*. In parallel, the draft genome assembly of the wheat genome released by Brenchley and collaborators [30] was mined for *TaMET1* genes. However, no full-length sequences corresponding to *TaMET1* genes were present in the dataset. Taking advantage of the recent release of sequence surveys from the International Wheat Genome Sequencing Consortium (IWGSC) (http://www.wheatgenome.org/) that were produced from sorted chromosome arms [18], BLASTn analyses against each chromosome arm were performed using rice and brachypodium *MET1* genes. Eventually nine *MET1* copies were identified and assigned to chromosomes 2A/2B/2D, 5A/5B/5D and 7A/7B/7D. For simplicity, homoeologous chromosomes A, B and D will be collectively referred to as a "homoeologous group" hereafter. Intron and exon junctions were defined for the nine *TaMET1* genes according to rice and brachypodium *MET1* genes and subsequently validated by RNA-seq analysis (see below). Protein domains were then predicted using the Pfam database. Three major protein domains were identified that include DNMT1-RFD (Cytosine specific DNA methyltransferase replication foci domain), BAH (Bromo-Adjacent Homology) and the DNA methyltransferase (C5 cytosine specific DNA methylase) domain (Figure 1). Comparison with *MET1* genes from rice orthologs showed an overall conservation of the *TaMET1* genes (Figure 1). Coding sequence analyses revealed that the *TaMET1* genes of chromosome 5A and 5D display deletions and premature stop codons (Figure 1) and if expressed produce truncated proteins missing the DNA methyltransferase domain. *TaMET-5A1* and *5D1* may be considered as pseudogenes, while all the remaining genes are expected to be functional.

### *TaMET1* loci originated from two successive duplication events

In a first attempt to understand the genome rearrangements, which have led to the nine *TaMET1* genes present in the *T. aestivum* genome, we retraced the phylogenetic history of *TaMET1* genes using *MET1* orthologs from monocotyledonous and dicotyledonous species. Two distinct copies of *MET1* (*i.e.* two distinct paralogs) are usually found in monocots such as rice, sorghum and brachypodium species. Phylogenetic analysis suggests that homoeologous *TaMET1* genes from group 2 are orthologous to *OsMET1b* on chromosome 7 while homoeologous *TaMET1* genes from group 5 and 7 are orthologous to *OsMET1a* on chromosome 3 (Figure 2A). Hereafter, these two phylogenetic groups are called MET-1a and MET-1b lineages in respect to the *MET1* genes from rice. The phylogenetic tree suggests that a first duplication event occurred early during monocot speciation resulting in the MET-1a and the MET-1b lineages (Figure 2A). Since these two copies of *MET1* are common to all grass species, the first *TaMET1* duplication is likely to be a consequence of the WGD that took place in all grasses and occurred about 56–73 MYa. The second duplication is shared only within the *Triticea* tribe (barley and wheat in our phylogenetic tree). Since wheat diverged from brachypodium 32–39 MYa and from barley 10–13 MYa [1,9], this second duplication probably occurred between 32 and 13 MYa. In order to understand if this duplication was the result of segmental or single gene duplication, syntenic relationships between regions surrounding the *TaMET1* genes from chromosome 5 and 7 and their orthologous loci in rice and brachypodium were investigated. For chromosome 5A, 5B and 5D, up to 80% of the genes were conserved, whereas only 10-15% were for group 7, suggesting that a single gene duplication occurred (Figure 2B). This hypothesis is consistent with the evolutionary model of grass genomes [11,31].

In order to date the duplication event leading to group 5 paralogs, BLASTn analyses were conducted between hexaploid wheat (*Triticum aestivum*), diploid wheat species (*Triticum urartu, Aegilops tauschii*) and barley (*Hordeum vulgare*). *Triticum urartu* (genome A ancestor) shares the same deletion with *TaMET-5A1* while *Aegilops tauschii* (genome D ancestor) and *TaMET-5D1* do not (Figure 2C). It can then be suggested that 5A was already in the process of pseudogenization before polyploidization while 5D pseudogenization occurred in the course of, or after, polyploidization. Consistent with this hypothesis, 5A displays a more pronounced gene structure alteration than 5D (large deletion and numerous stop codons; see also Figure 1).

### *TaMET1* genes display distinctive evolution rates

We then investigated the putative functional differences between the nine *TaMET1* genes by evaluation of the evolution rate, which is a good indicator for the biological function of a given gene [32]. We chose the codon substitution model to estimate the rate of synonymous

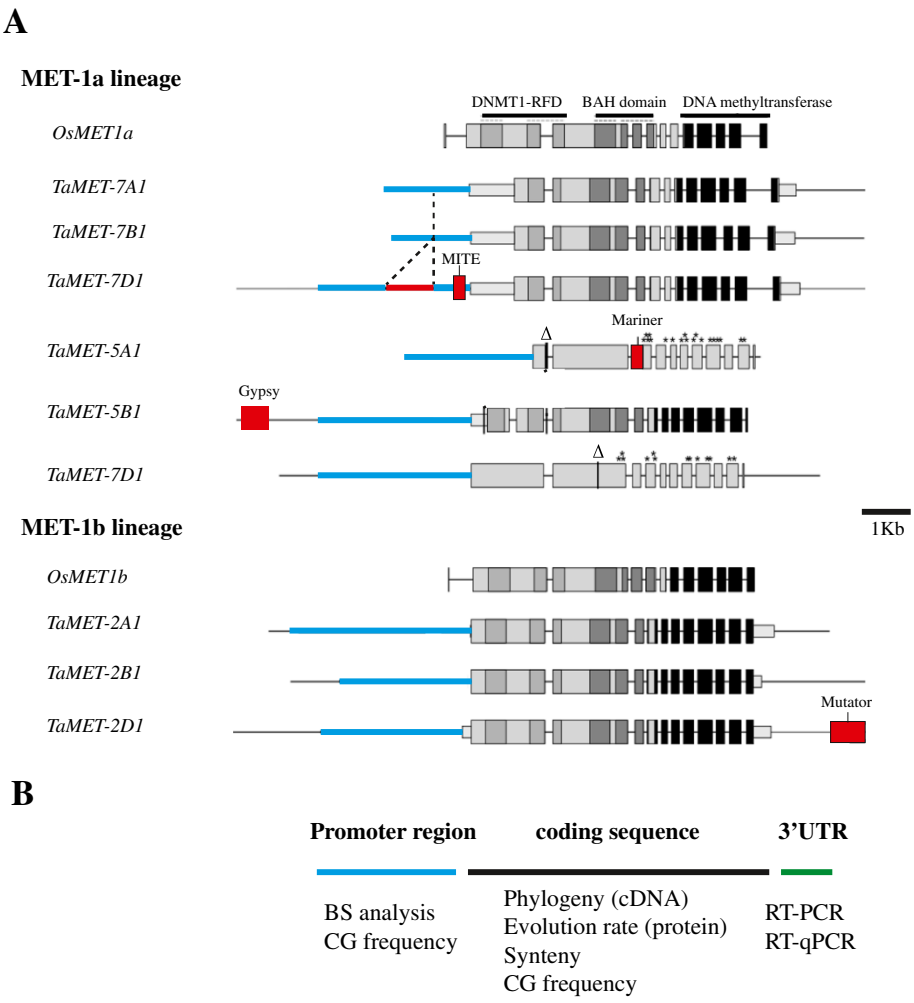



**Figure 1 Gene structures of *TaMET1* genes. A)** The two *MET1* genes from rice (*OsMET1a*; [GenBank: Os03g58400] and *OsMET1b*; [GenBank: Os07g08500]) were used as a reference to define two MET1 lineages (hereafter called MET-1a and MET-1b lineages). *OsMET1a* and *OsMET1b* are shown at the top of each lineage. Gene structures and splice junctions organize *TaMET1* genes in 11 exons (black and grey boxes) and 10 introns (horizontal lines). The three distinct protein domains identified are indicated at the top of the figure: DNMT1-RFD (light grey), BAH-domain (dark grey) and DNA methyltransferase (black). *TaMET1* genes from chromosome 2 and 7 as well as chromosome 5B are predicted to yield a full-length MET1 protein while 5A and 5D contain stop codons (*) and deletions (Δ). 5′ and 3′UTRs are indicated as smaller boxes. Potential promoter regions were defined as the ~2-3 kb region upstream of the coding sequences and are indicated as a thicker blue line. Unknown sequence insertion (*TaMET-7D1*) is indicated as a red line and transposable element insertion of Stowaway, Gypsy, Mariner and Mutator are indicated as red boxes. **B)** Methods used in this study.

(dS) and non-synonymous (dN) substitutions and computed the dN/dS ratio as evolution rate (ω) [33]. In this model, for genes with a significant biological function undergoing purifying selection non-synonymous mutations are expected to be kept at a low level whereas synonymous mutations accumulate randomly according to the neutral theory of evolution [33].

As a first approach, pair-wise divergences were investigated between *TaMET1* genes and *MET1* genes from fully sequenced monocot species (*i.e.* divergence between orthologous pairs). Mean values for A, B and D homoeologs were then calculated per homoeologous group of chromosomes (group 2, 5 and 7) and are displayed in Figure 3A. Consistent with the neutral theory of evolution, dS rates were not significantly different between the three homoeologous groups. However significant differences were observed for dN and ω indicating a lower rate of evolution for homoeologous group 2 which belongs to the MET-1b lineage. Homoeologous group 7 is evolving at an intermediate evolution rate compared to group 2 and group 5 but does not display any deleterious mutations within the coding sequences (see also Figure 1). As expected for pseudogenes, higher dN and ω values were found for *TaMET1* at homoeologous group 5.

Secondly, various hypotheses concerning evolution rates were then tested and Likelihood Rate Tests (LRT)

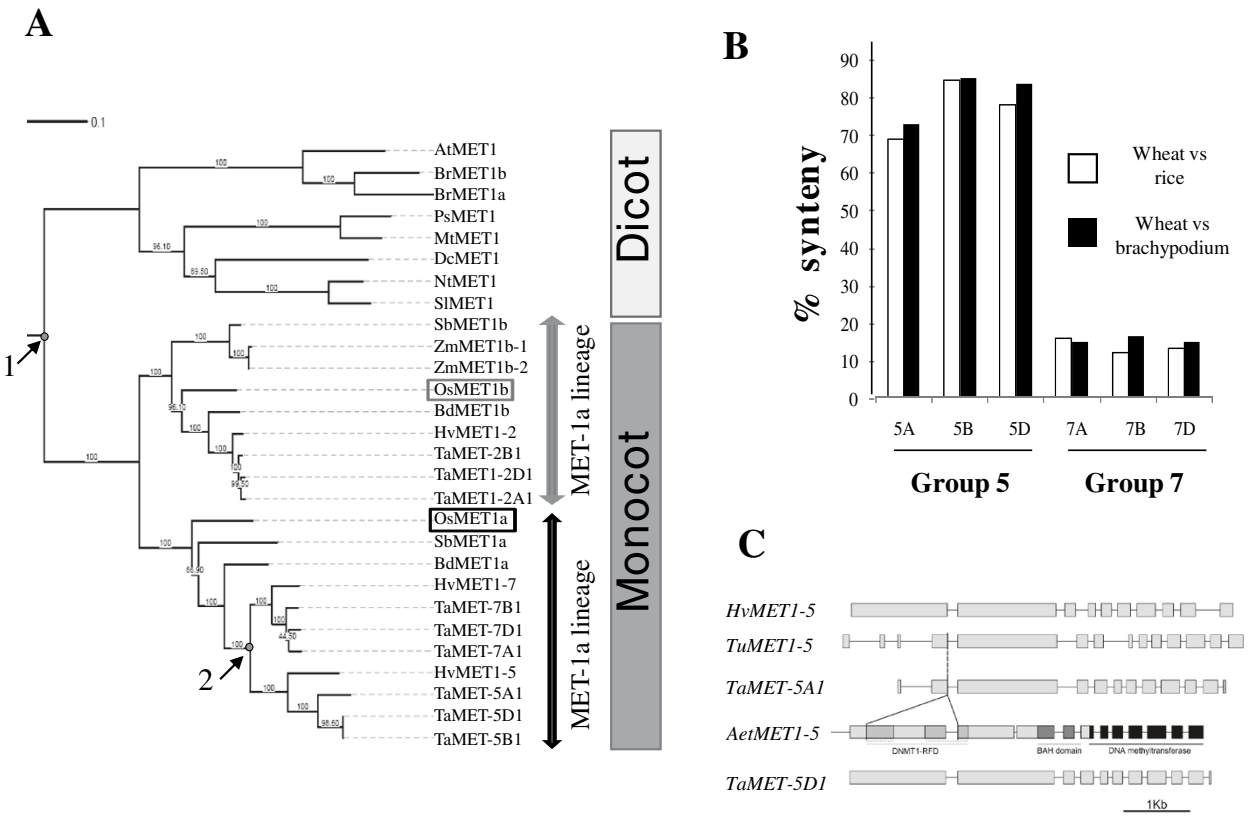



**Figure 2 *MET1* phylogenetic trees in flowering plants and syntenic conservation. A)** Maximum likelihood (ML) phylogenetic tree. The ML tree was inferred using cDNA sequences from dicot species including *Arabidopsis thaliana* (*AtMET1*: [TAIR:At5g49160]), *Brassica rapa* (*BrMET1a* and *BrMET1b* respectively [GenBank: AB251938 and AB25937]), *Pisum sativum* (*PsMET1*: [GenBank: AF034419]), *Medicago truncatula*: (*MtMET1*: [Phytozome: Medtr6g065580]), *Daucus carota* (*DcMET1*: [GenBank: AF007807]), *Nicotiana tabacum* (*NtMET1*: [Genbank: AB030726]), *Solanum lycopersicum* (*SlMET1*: [GenBank: AJ002140]), and monocot species including *Sorghum bicolor* (*SbMET1a* and *SbMET1b* respectively [Phytozome: Sb01g005084 and Sb02g004680]), *Zea mays* (*ZmMET1b-1* and *ZmMET1b-2* respectively [Phytozome: GRMZM2G333916 and GRMZM2G334041]), *Oriza sativa* (*OsMET1a* and *OsMET1b* respectively: [Genbank AB362510 and AB362511]), *Brachypodium distachyon (BdMET1a* and *BdMET1b* respectively [Phytozome: Bradi1g05380 and Bradi1g55287]), *Hordeum vulgare* (HvMET1-2 and HvMET1-7 respectively*:* [Ensembl Genomes: MLOC_61904.6 and: MLOC_10988.2], *HvMET1-5* predicted from [GenBank: CAJW010043285]) and *Triticum aestivum*: *TaMET-2A1*, *2B1*, *2D1*, *5A1*, *5B1*, *5D1*, *7A1*, *7B1* and *7D1* (this study from IWGSC contigs). Numbers above branches indicate bootstrap values. Two monophyletic groups were defined and called the MET-1a and MET-1b lineages according to the *OsMET1a* and *OsMET1b* genes (indicated in boxes). The two gene duplication events observed in wheat and barley are indicated by an arrow. **B)** Micro-synteny analysis. Micro-synteny was established by BLASTn analysis between genes from brachypodium, rice and IWGSC sequences from wheat chromosomes as described in Methods. Syntenic conservation between rice (in white) and brachypodium (in black) correspond to the number of conserved genes in wheat/number of genes in the syntenic region from brachypodium or rice expressed as percentage. **C)** Gene structures of *MET1* genes at chromosome 5. BLASTn analysis was used to align *MET1* genes from *Hordeum vulgare* (*HvMET1-5*: [GenBank: CAJW010043285]), *Triticum urartu* (*TuMET1-5* [Ensembl Plants: scaffold15783]), *Aegilops tauschii* (*AetMET1-5* [Ensembl Plants: scaffold164515]) and *Triticum aestivum* (*TaMET-5A1* and *TaMET-5D1*) (this study).

were computed. Eleven hypotheses were evaluated to test an increased evolution rate at various branch points in the phylogenetic tree (Additional file 1). Evolution rates are summarized in Figure 3B. The results support the existence of three evolution rates (indicated as ω0, ω1 and ω2 in Figure 3B) consistent with the two duplication events and the pair-wise analysis performed previously (Figure 3A). After gene duplication, long-term changes were observed in our phylogenetic tree. ω0, ω1 and ω2 evolution rates indicate that negative selection occurs in the MET-1b lineage, which has the smallest evolution rate (ω0 = 0.12193) suggesting its functional role in monocots. Following the first duplication event, a two fold increase in evolution rate (ω1 = 0.21988) is observed in the Met-1a lineage except for barley chromosome 5 and wheat homoeologous group 5 for which a fourfold increase (ω2 = 0.48706) is observed.

Altogether, evolution rate analyses indicate that *TaMET1* homoeologous genes of group 2 are submitted to stronger purifying selection and are evolving at a slower rate suggesting their predominant role in DNA methylation maintenance in hexaploid wheat. Following the second duplication event, asymmetric acceleration of the evolution rate is observed between homoeologous

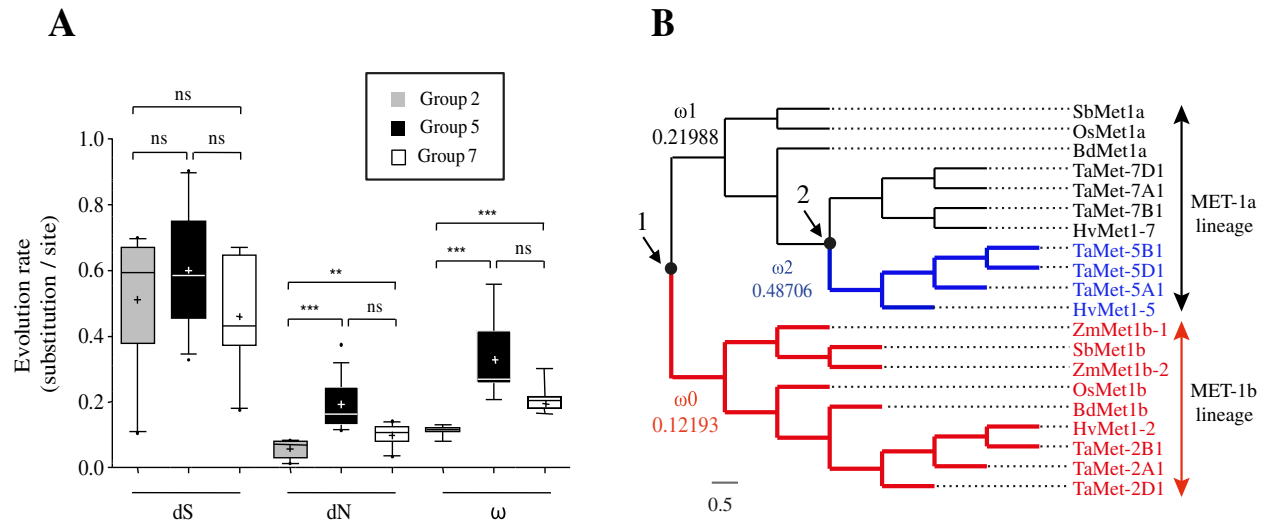



**Figure 3 Evolution rate analyses of *TaMET1* genes. A)** Pair-wise comparisons. Synonymous (dS), non-synonymous (dN) and evolution rate (ω) are expressed in substitution/site and were performed using *Codeml*. *TaMET1* from wheat homoeologous group 2 (grey boxes) were compared to *Brachypodium distachyon* [Phytozome: Bradi1g05380 and Bradi1g55287], *Sorghum bicolor* [Phytozome: Sb01g005084 and Sb02g004680], *Oriza sativa* [Genbank: AB362510 and AB362511], *Zea mays* [Phytozome: GRMZM2G333916 and GRMZM2G334041], *Hordeum vulgare* [Ensembl Genomes: MLOC_61904.6 and: MLOC_10988.2 and GenBank: CAJW010043285] and *Triticum aestivum* (*TaMET-2A1*, *2B1*, *2D1*, *5A1*, *5B1*, *5D1*, *7A1*, *7B1* and *7D1*) (accession numbers in Additional file 8). Mean values of dS, dN and ω (with ω = dN/dS) were then computed for each homoeologous group. Whiskers represent the 10-90% range of mean values, boxes represent interquartile distances, the horizontal line across whiskers represents the median, and "+" the mean values. Kruskal Wallis non-parametric tests were applied to determine significant differences between mean values (*: $P < 0.05$; **: $P < 0.01$; ***: $P < 0.001$). **B)** PAML branch model. Tree topology was defined by a protein alignment using the same monocot species as in A). The two monophyletic groups MET-1a and MET-1b are indicated at the right. Three distinct evolution rates ω0 (red branches), ω1 (black branches) and ω2 (blue branches) are indicated as well as the two gene duplication events described in Figure 1 (arrow).

group 5 and 7 leading eventually to the formation of pseudogenes within group 5 that accumulated deleterious mutations within their coding sequences.

### Expression of *TaMET1* genes is mainly driven by homoeologous group 2

The above analysis of evolution rates suggests that homoeologous genes from group 2 are under purifying selection. As it is well documented that expression patterns influence non-synonymous substitution [34], expression levels and profiles of the nine *TaMET1* genes were investigated using RNA seq data from five different organs at three developmental stages each. Expression levels of *TaMET1* homoeologous genes from group 2 were found to be 10 to 40 times higher than the *TaMET-5B1* and *TaMET* group 7 ones. For *TaMET-5A1* and *5D1*, no significant expression was detected in any of the 15 conditions (Figure 4A). Homoeologous group 2 were expressed in most tissues at nearly all developmental stages, named according to the Zadoks (Z) scale [35], but with highest expression levels at Z30 in the stem and Z32 in the spike. *MET1* expression levels in other species peak in proliferating cells such as in meristems and in reproductive organs [27,28,36]. In wheat we observed *TaMET1* expression at early developmental stages especially during early stem extension (Z30-Z32) when wheat is switching from the vegetative to reproductive phase. At that stage the spike tissue is proliferating requiring active replication during which DNA methylation maintenance should occur. Similarly, homoeologous group 7 were found to be expressed in almost all conditions but at a very low level compared to group 2 genes. A similar situation was observed in rice where *OsMET1a* is 10–12 times less expressed than *OsMET1b* [37]. For homoeologous group 5, only 5B is expressed at low level in grain (Figure 4A).

RNA-seq-based expression profiles were subsequently confirmed by RT-PCR. Various primer pairs were designed at the 3′UTR. Semi-quantitative and quantitative analyses confirmed the expression of *TaMET1* from homoeologous group 2 (Figures 4B and 4C) but transcripts were hardly or not detectable for group 5 and 7 (Figure 4B). Expression levels for 2A, 2B and 2D homoeologs were investigated by RT-qPCR but did not show strong differences, although 2B was found to be slightly more expressed (Figure 4C). Thus *TaMET-2A1*, *2B1* and *2D1* are expressed in an additive mode.

Recent analyses at the whole genome level indicated that housekeeping genes in wheat are enriched at pericentric positions while genes expressed with tissue-specific patterns and pseudogenes are usually found at more sub-telomeric positions [38]. To check whether there is a

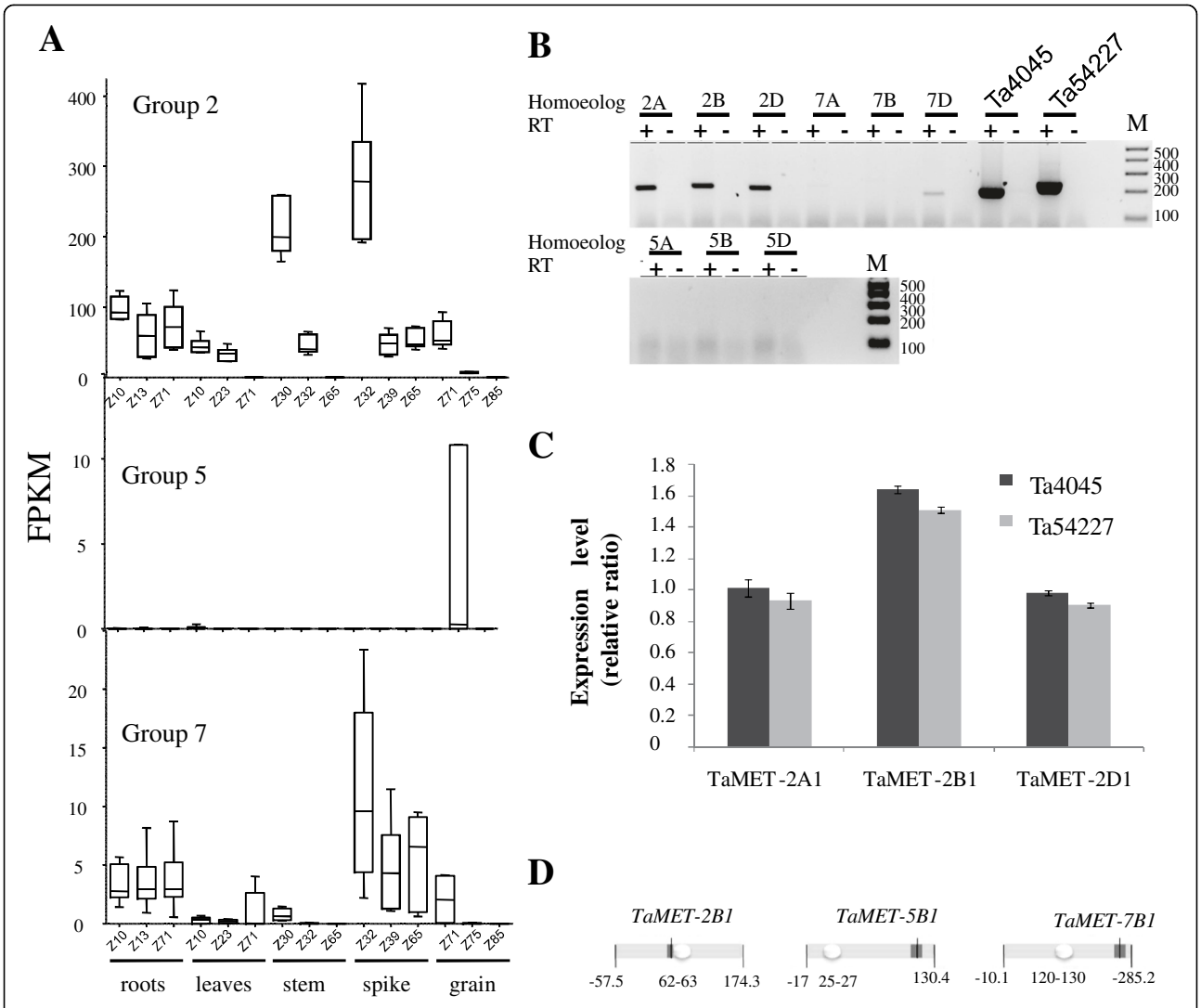



**Figure 4 Expression profiles of wheat *TaMET1*. A)** Wheat RNA-seq. Data are expressed in Fragment Reads per Kilobase of Exon Model (FPKM) for each homoeologous group. FPKM were computed according to the following formula: FPKM = $10^9$ x (C/NL) where C is the number of mappable reads on a feature, N is the total number of reads in the experiment and L (length) is the sum of exonic sequences in base pairs. Genes were considered to be expressed only for FPKM values >0.15. Samples covering various tissues (root, leave, stem, spike and grain) and developmental stages indicated as Zadoks (Z) scales are indicated at the bottom. Expression levels from *TaMET1* genes of a given homoeologous group were used to compute mean values. *TaMET1* from group 2 are expressed at higher levels in comparison to the other *MET1* copies and display two main peaks of expression in the stem (Z30) and spike (Z32). **B)** RT-PCR gel analysis. RT-PCR with (+) or without (–) reverse transcriptase was performed for all *TaMET1* genes. Control PCR reactions were performed for two constitutively expressed genes corresponding to Ta4045 and Ta54227 selected according to Paolacci *et al.* 2009. **C)** Quantitative RT PCR. Reactions were performed for *TaMET-2A1*, *2B1* and *2D1*. Expression is relative to Ta4045 (dark grey) and Ta54227 (light grey). **D)** Schematic representation of *TaMET1* loci. Position of *MET1* genes on chromosome maps of 2B, 5B and 7A (chromosome arms in grey and centromere as white ellipse, not to scale). Top and bottom positions on ITMI reference map are indicated in cM.

correlation between the observed gene expression differences and the physical position of *TaMET1* copies on the chromosomes, we genetically mapped *TaMET1* loci using 57 SNPs identified in the course of our sequence capture experiments (see Methods). Out of the 57 SNPs, 18 produced high quality results that led to the genetic mapping of five out of nine *TaMET1* genes, namely *TaMET-2B1*, *5A1*, *5B1*, *7A1* and *7B1*. As positions of homoeologous copies were consistent for groups 5 and 7, we extrapolated the position of all *TaMET1* genes from these five copies. Homoeologous group 2 were found to be located in the pericentromeric regions of the short arm of chromosomes 2 whereas group 5 and 7 were mapped to subtelomeric positions of the long arms of chromosomes 5 and 7 respectively (Figure 4D and Additional file 2).

Thus *MET1* expression is mainly driven by homoeologous group 2 indicating specific mechanisms to keep a predominant expression of homoeologous group 2 over groups 5 and 7. This observation resembles a phenomenon observed after *METI* gene duplication in Arabidopsis where *METI* transcripts accumulate to 10,000 fold higher levels than those of the duplicated *METIIa* and *b,* while *METIII* is considered to be a pseudogene [39]. Expression of a specific member of a given gene family is referred to as predominance [40] or transcriptional dominance [16] and in our case occurs for *TaMET1* genes at homoeologous group 2. The pericentric position of group 2 genes is consistent with the conclusions drawn from a recent large scale analysis indicating that genes expressed in most tissues are located in more proximal position than those displaying tissue-specific expression patterns [38]. Thus expression studies reinforce the idea that *MET1* homoeologous group 2 genes might provide methyltransferase activity.

### CG-rich isochores appear at *TaMET1* promoters and exhibit high DNA methylation

While low, or absent expression of specific *TaMET1* genes might be explained by several factors including genetic mutations or insertion of transposable elements, epigenetic marks at promoter regions are good candidates to explain differences in gene expression [23,24,41,42]. Among these, cytosine methylation that occurs in CG sequence contexts has been shown to modulate gene expression in plants [23,24,41,42]. To investigate the potential role of DNA methylation in the regulation of the *MET1* genes, *MET1* coding sequences as well as putative promoters were scanned for potential methylation sites in CG, CHG and CHH sequence contexts.

The putative promoters of the nine genes were defined as ~2-3 kb regions upstream of the coding sequence depending upon the availability of the genomic sequences (Figure 1). Comparisons between upstream and coding sequences for potential methylation sites in CHG and CHH contexts revealed similar amounts of CHG and CHH sites for all nine genes (data not shown). In contrast, cytosines in the CG context were enriched at potential promoter regions of homoeologous group 5 (4.4 fold) and group 7 (5.5 fold) compared to group 2 putative promoters regions (Figures 5A and 5B). This result was unexpected because CG-rich regions (also known as CG-rich isochores), although already described in Arabidopsis genes, were shown to be mainly located in introns [43].

As CG-rich isochores at *TaMET1* upstream regions could be the result of new insertions of CG-rich DNA sequences, we looked for such events. Indeed, two DNA insertions of 786 and 122 bp overlapping with CG-rich isochores were observed for the *TaMET-7D1* upstream region (Additional file 3). Both insertions were already present within the ancestral D genome (*Aegilops tauschii*) suggesting their integration prior to polyploidization (Additional file 3). BLASTn analysis against the TREP database indicated a short but significant homology with a *stowaway* Miniature Inverted Repeat (MITE) for the 122 bp insertion while no significant homology was detected for the larger 786 bp insertion. BLASTn against TREP performed with the five remaining upstream regions (7A, 7B, 5A, 5B and 5D) failed to detect any transposable elements as shown in Figure 1. Instead of a new large DNA insertion enriched in CG observed at 7D, the CG-rich isochores are more dispersed along the 5A, 5B, 5D, 7A and 7B putative promoter regions (Figure 5A). This may argue in favor of a progressive CG accumulation in the course of evolution.

To determine whether these regions enriched in cytosine residues on homoeologous group 5 and 7 are indeed methylated, we performed bisulfite sequencing. We designed bisulfite primers in a way to simultaneously amplify all three homoeologous copies that we subsequently discriminated upon sequencing. Consistent with our expression studies, putative promoter regions from homoeologous groups 5 and 7 display DNA methylation in CG sequence contexts. Homoeologous group 5 also displays significant CHG methylation (Figure 5C and Additional files 4 and 5). Among all the analyzed putative promoter regions, the highest DNA methylation levels overlap with the 786 bp insertion specific to 7D (Figure 5C and Additional file 5).

Taken together, our results suggest that the presence of CG-rich isochores in the putative promoters of group 5 and 7 *TaMET1* homoeologous genes may be due to a progressive and dispersed CG-enrichment as well as to an insertion-mediated CG-enrichment, at least for the 7D copy. In addition, the high methylation levels observed in the promoter regions of the two low-expressed homoeologous groups may suggest the existence of an autoregulatory loop controlling *MET1* gene expression.

## Discussion

Bread wheat is a plant species with a large genome of about 17 Gb containing up to 80% of repetitive sequences. Much attention has been focused recently to understand how this genome, highly enriched in repetitive sequences, controls its transposable element fraction, which will otherwise lead to genome instability. One such mechanism is likely to involve DNA methylation in the CG context, which is maintained by MET1. It is therefore of importance to understand how MET1 expression is regulated in an organism with a complex hexaploid genome. In the course of our work, we observed that *TaMET1* genes contain a record of many evolutionary events, which have occurred prior and after the emergence of bread wheat.

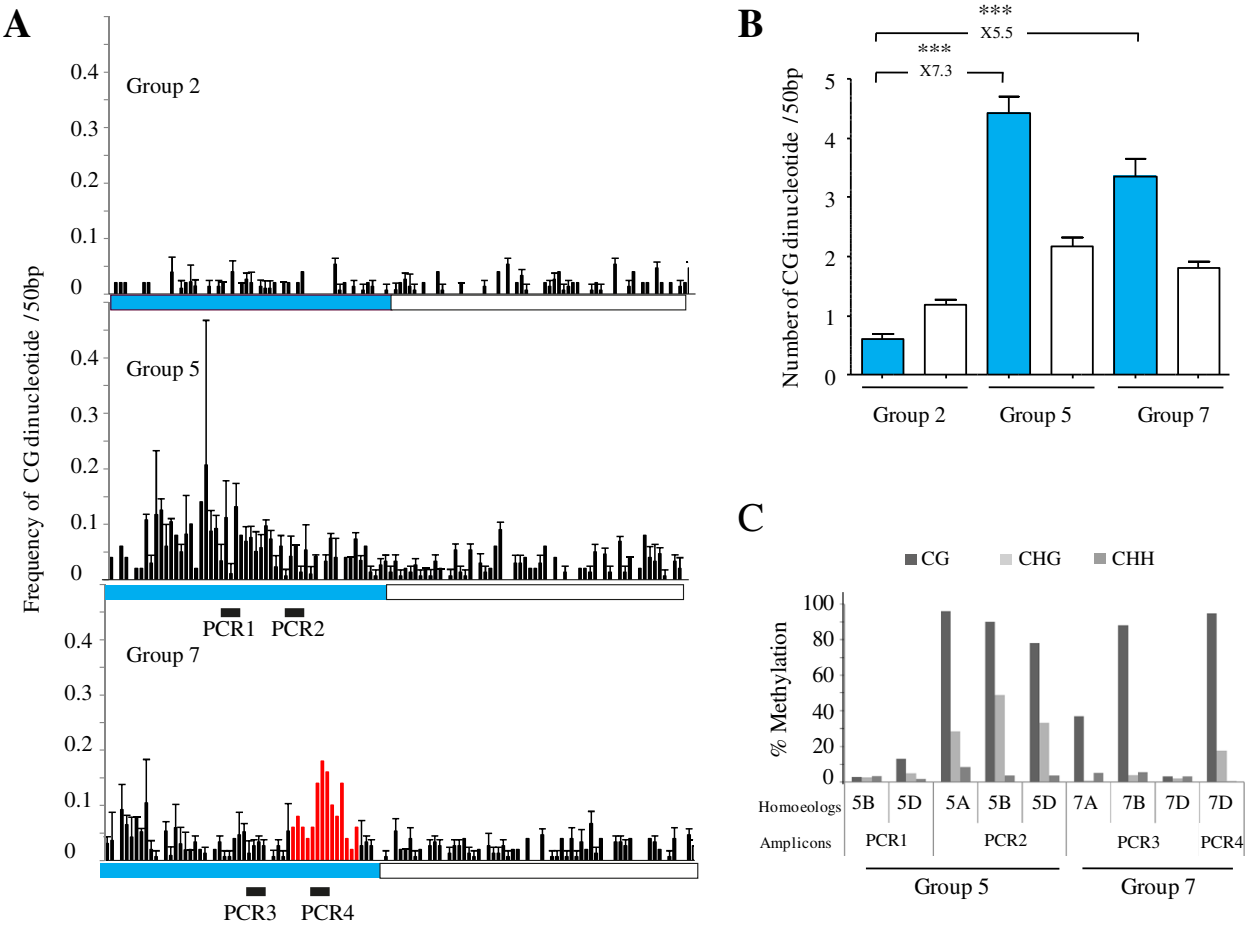



**Figure 5 CG enrichment and methylation at potential promoters of *TaMET1* genes from homoeologous group 5 and 7. A)** Frequency of CG dinucleotides. Frequencies were computed every 50 bp and are shown for each homoeologous group. Putative promoter region and coding sequence are delimited by respectively a blue and white box. Black bars numbered from PCR1 to PCR4 highlight the four regions studied by bisulfite sequencing and are indicated above the graphs. Region 4 is specific to the 7D homoeolog. Arrows indicate the putative transcription start site according to the RNA-seq data. **B)** Mean values of CG dinucleotides. Mean values of the number of CG dinucleotides of the three homoeologs (A, B and D) for a given homoeologous group (2, 5 and 7) in the putative promoter (blue) and coding (white) sequence regions. Differences between groups 5 and 7 putative promoter regions and group 2 are indicated above the histogram. Statistical significance was confirmed with a Kruskal Wallis non parametric tests with *: $P < 0.05$; **: $P < 0.01$; ***: $P < 0.001$. **C)** DNA methylation profiles as determined by bisulfite sequencing. Percentages of methylated cytosines of the four amplicons (PCR1 to 4) displayed in Figure 6A were determined after bisulfite sequencing. Percentages of methylation were recorded at each cytosine position and were used to compute a mean value for each amplicon in the CG, CHG or CHH sequence contexts.

We identified nine copies of *TaMET1* organized in three homoeologous groups at chromosomes 2, 5 and 7. At the chromosomal level, segments bearing *TaMET1* originated from two duplication events. Phylogenetic and micro-synteny confirmed that chromosome 2 and 5 paralogs originated from a WGD about 50–70 MYa in the ancestor of grass species. Then the chromosome 7 paralog emerged from a more recent gene duplication about 13–32 MYa in the *Triticea* tribe. Our analysis of the evolution rate revealed functional differences between the nine *TaMET1* genes. The MET-1b lineage (homoeologous group 2) was shown to display a lower evolution rate than the MET-1a lineage (homoeologous group 5 and 7). Lower evolution rate is observed for genes with biological function and this is best explained by purifying selection, which counter selects deleterious mutations [34,44]. Functional significance of homoeologous group 2 genes was reinforced by our observations of expression levels and DNA methylation. Low evolution rate in the MET-1b lineage matches with a predominant expression of homoeologous group 2 over group 5 and 7. Predominant expression of one member of the *MET1* gene family was already observed in other species such as Arabidopsis [39] and rice [37] suggesting that MET1 expression level and pattern needs to be carefully controlled. Interestingly, we mapped *TaMET1* homoeologous group 2 to peri-centric (proximal) position

while group 5 and 7 were located at more sub-telomeric (distal) regions. Recent large scale analyses in wheat suggested that distal regions are more dynamic, displaying higher level of recombination and accumulate more pseudogenes and gene duplications than proximal peri-centric regions [38]. Furthermore, genes at distal position display more tissue specific expression than those at more proximal position. It is then tempting to hypothesise that a distal chromosome position may have a direct influence on expression leading as a consequence to the predominance of the more proximal genes as observed in our case for homoeologous group 2. Homoeologous group 2 did not show any differences in gene expression among the three homoeologs. Consistent with the whole genome analyses was the fact that genome-wide transcriptional dominance of an individual subgenome (A, B or D) was not observed [18]. Besides its position along the chromosome, our data indicated that DNA methylation observed in the promoter region of homoeologous group 5 and 7 may have contributed to their transcriptional repression and may have favored an increased evolution rate at *TaMET-5A1* and *5D1* leading to the accumulation of deleterious mutations, a process known as pseudogenization [45,46]. Interestingly, distinctions can be made between group 5 homoeologs: 5A already accumulated large deletions and numerous stop codons before polyploidization, while stop codons occurred at 5D after polyploidization but are absent at 5B which however displays an elevated level of non-synonymous mutations and is expressed only in grains. Pseudogenes are usually rapidly eliminated and the fact that *TaMET-5A1* and *5D1* pseudogenes are maintained suggests that pseudogenization may not be fully achieved or that these genes contribute in a significant but yet unknown manner to TaMET1 activity.

Our data support a functional role of DNA methylation in the initiation or the maintenance of gene silencing at specific *TaMET1* genes. Considering that the chromosome 2 paralog is the ancestral locus and shows low occurrence of potential CG methylation sites, the observed CG-rich isochores at chromosome 5 and 7 paralogs associated with DNA methylation imply CG-enrichment at these putative promoter regions. CG-enrichment was unexpected as usually CG dinucleotides are under-represented due to 5-methylcytosine deamination [43]. At the moment we can only speculate about their possible origin. First, GC-rich and GC-poor isochores are known to occur in animals and several hypotheses have been proposed to explain their emergence [47]. Among them the GC-biaised gene conversion (gBGC) has been proposed as one of the main driving forces in the evolution of nucleotide composition. gBGC favors GC over AT bases in alleles during mismatch repair following heteroduplex formation in the course of meiosis. gBGC results from Base Excision Repair (BER) and involves a DNA glycosylase that specifically removes thymine in DNA heteroduplexes. Secondly, animal genomes display unmethylated CG-rich elements known as CpG islands (CGIs). CGIs are defined as DNA sequences of a few hundred base pairs, with high CG occurrence, high G + C frequency and are involved in the regulation of gene expression [48]. CGIs have been divided into start and non-start CGIs. Non-start CGI are the most abundant and best explained by insertion of repeated sequences such as transposable elements (in the human genome 79% are due to *Alus*) while start CGIs located at the transcription start sites are only poorly explained by transposable element insertion (in the human genome 5,6% are due to *Alus*) [49]. Interestingly, Suzuki et al., [50] also proposed gBGC as one of the possible mechanisms to explain the emergence of start CGIs. Recently, it was suggested that gBGC occurs in plants [51]. gBGC can be considered as one of the possible mechanisms explaining the emergence of CG-rich isochores at *TaMET1* putative promoter regions. Indeed, it may be an attractive mechanism to explain the progressive CG enrichment we observed at *TaMET1* upstream regions especially at homoeologous group 5 and 7 located at distal chromosome positions where higher recombination rates have been reported [38,52]. Furthermore, the *MET-7D1* copy would have also undergone insertion of CG-rich DNA fragments in a mechanism very reminiscent to the one observed for non-start CGIs, arguing for shared evolutionary mechanisms between animal and plants toward the emergence of CG-rich isochores.

Once CG-rich isochores appeared, they can be methylated in order to silence gene expression. Although CGIs were not described in plant promoters, “dense CG methylation clusters” have been reported and are proposed to silence cryptic promoters within the coding sequence [43]. Silencing of these cryptic promoters is established first through the RNA-directed DNA Methylation (RdDM) pathway and results in methylation at cytosine residues at CG, CHG and CHH sequence contexts. Once methylation is set up, only methylation in the CG context, which does not rely on siRNA production, can be maintained in the course of evolution leading to high methylation only in CG sequence contexts [43]. If such a mechanism occurred within the putative promoter region of *TaMET1* genes, it can explain how homoeologous group 7 became progressively repressed.

Given the correlation between DNA methylation in promoter regions and gene silencing [23,24], we suggest that DNA methylation may be part of a possible autoregulatory mechanism among *TaMET1* genes. In this model, MET1 mainly encoded by homoeologous group 2 regulates group 7 gene expression through CG DNA methylation maintenance. CG methylation at homoeologous group 7 may be alleviated in specific organs, developmental stages or upon changing environmental conditions. However possible roles for the homoeologous group 7

(MET1-a lineage) is challenged by recent data collected in rice indicating that the main MET1 function is ensured by *Met1b* and not *Met1a*. Indeed, RNAi against *Met1a* does not significantly affect plant development while a *met1b* null mutant is lethal [28,53].

## Conclusions

From our data, we propose a chronology (Figure 6) of the genomic events observed at *TaMET1* genes, which include WGD, gene duplication, expression predominance of homoeologous group 2, CG-rich isochores emergence, DNA methylation and pseudogenization. The unexpectedly rich evolution history observed at *TaMET1* makes these loci a very attractive model to study further gene evolutionary mechanisms occurring in hexaploid wheat. Increased copy number finally leads to *TaMET1* silencing at homoeologous group 5 and 7 (the MET-1a lineage), keeping genes of group 2 (the MET-1b lineage) in an active state. We hypothesize that CG methylation was used as a mean to control gene expression in the MET-1a lineage favoring low expression at homoeologous group 7 and pseudogenization at group 5. For the latter the different evolutionary stages are still observed between homoeologs. CG methylation might be required to limit homoeologous group 7 transcription using CG-rich isochores, which are the target of CG DNA methylation. At that point, we can only speculate about the possible involvement of methylation in limiting homoeologous group 7 expression in tissues or developmental stages where group 2 is expressed, maybe leading to tissue-specific expression patterns of group 7 genes and their subfunctionalization.

Taken together, our data suggest that DNA methylation at *TaMET1* loci can act as an epigenetic determinant required to drive genome evolution.

## Methods

### Analysis of IWGSC sequence surveys

Access to the IWGSC physical map sequences from hexaploid wheat *cv Chinese Spring* has been established by the URGI (http://urgi.versailles.inra.fr/) on the behalf of IWGSC. BLASTn analyses were performed at the URGI database. Identification of transposable elements in selected IWGSC contigs was performed by BLASTn analyses against the TREP database at http://wheat.pw.usda.gov/ITMI/Repeats/. CG, CHG and CHH profiles (where H is A, C or T) were detected using an in house Perl script available upon request. Gene structures were predicted by cDNA alignment against genomic sequences

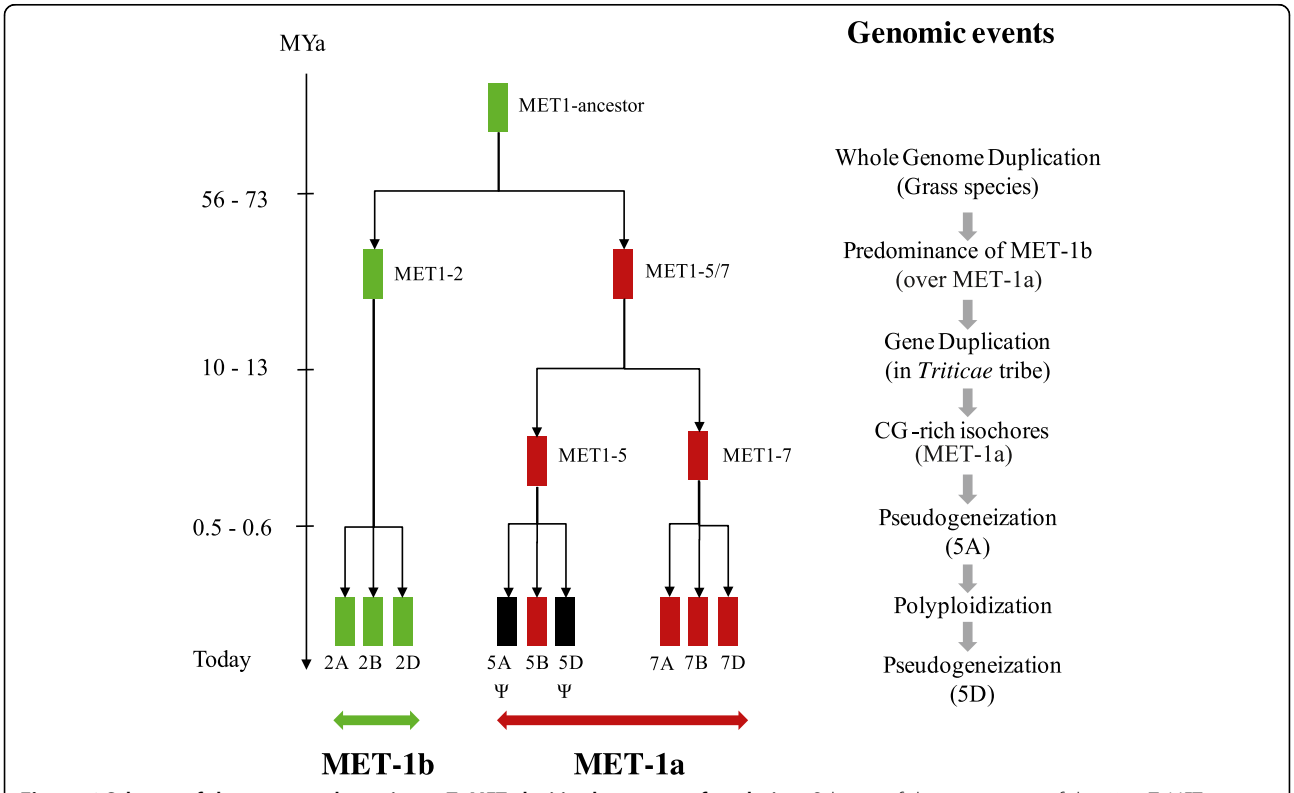



**Figure 6 Scheme of the genome dynamics at *TaMET1* loci in the course of evolution.** Scheme of the emergence of the nine *TaMET1* genes is given at the left. Evolution time is given from the top to the bottom in MYa. The MET-1a lineage is expressed (green) and evolved at a low evolution rate while genes in the MET-1b lineage are repressed (red) or evolved as pseudogenes (Ψ, black boxes). The hypothetic succession of genomic events occurring at or including *TaMET1* is proposed at the right.

using *SIM4* (http://pbil.univ-lyon1.fr); protein domains are according to the Pfam database (http://pfam.sanger.ac.uk/). Gene structures were subsequently designed by *FancyGene* (http://bio.ieo.eu/fancygene/).

### Sequence capture

Sequence capture experiments [54,55] were designed to isolate and sequence DNA segments using probes synthesized on a microarray. Two runs of sequence capture were performed according to the NimbleGen Arrays User's Guide followed by 454 Optimized Sequence Capture method. Briefly, *MET1* specific probes ranging from 60 to 90 nucleotides were designed at a density of ~$1.10^6$ probes/Mb of sequence by Roche-Nimblegen from one *Triticum aestivum* public EST (accession number TA8302), one private EST (GPIC:S:720428) and from the two coding sequences from *Brachypodium distachyon* (Accession numbers [Phytozome: Bradi1g05380 and Bradi1g55290]). The absence of repeated sequences was verified using repeatMasker and TREP release 10 database. Genomic DNA from wheat elite lines cv *Brigadier*, *Alcedo*, *Renan* and *Recital* were used to build up four distinct genomic DNA libraries by nebulization with an average of fragment sizes of ~600 bp. Libraries were then hybridized onto capture arrays, captured DNA was eluted and amplified prior to 454 sequencing on a GS FLX Titanium platform according to the manufacturer. Overall, sequence captures yielded 8,184 reads specific to *MET1* and *de novo* assembly was subsequently performed using *gsAssembler* (Roche) with specific parameters set at 98% similarity and 20 bp overlap. Sequences were verified in the course of the project by BLASTn analysis against the IWGSC surveys and by PCR amplification on diploid and hexaploid species.

### Genetic mapping

Two mapping populations were used: recombinant Inbred Lines derived from a cross between *Triticum aestivum cv Renan* and *Recital* [56] and a doubled haploid population derived from a cross between *Triticum aestivum cv Brigadier* and *Alcedo* (Biogemma personal communication). DNA from all four elite lines was used in the sequence capture experiments and reads were grouped according to the four DNA origins. As a whole, 57 putative Single Nucleotide Polymorphisms (SNPs) were identified and genotyping was subsequently performed on genomic DNA from the two mapping populations using KASPar (KBioscience) fluorescent competitive allele-specific amplification. Primers were designed with *Primer picker* (KBioscience) and PCR amplifications were performed on a hydrocycler (LGC genomics), for 41 to 50 cycles at 57°C and then run onto a Genotyper (Applied Biosystem). The list of primers used to perform the genetic mapping is provided in Additional file 6. SNP mapping was performed on the two genetic maps using an in-house bioinformatic pipeline available at Biogemma. Genetic positions are given according to the *Renan* x *Recital* recombination map [56]. Physical positions are according to the names of the IWGSC contigs obtained by BLASTn analysis against the virtual map and are included within the virtual map designed by synteny.

### Phylogenetic reconstruction and substitution rate calculation

*TaMET1* coding sequences were used for phylogenetic reconstruction and substitution rate calculation. Selected sequences were first aligned with *MUSCLE* multiple sequence alignment [57] and then refined using *Gblocks* [58]. Maximum likelihood analysis was performed with *PhyML* using default parameters with 1,000 bootstraps [59]. Phylogenetic trees were drawn using *ITOL* [60]. Substitution rate studies were performed as follows: first, a new phylogenetic tree was built with the same species except that here, dicot species were not considered and the tree was based on protein sequences instead of cDNA. For *TaMET1*, genomic sequences were used to predict exonic sequences using *FGENESH* [61] and subsequently assembled into a predicted cDNA. Predicted cDNAs were validated in the course of this study by RNA-seq data. cDNAs were translated using *Transeq* and *Sixpack* from the *EMBOSS* package [62]. The phylogenetic tree was then built from predicted proteins as described above. ω (the ratio of nonsynonymous/synonymous substitution rates) was determined using *Codeml* from the *PaML* package [33]. A likelihood ratio test (LRT) was used to compare various hypothesis models in which ω values are expected to differ among branches, in comparison to a null hypothesis in which all the branches have similar ω. LRT values were then compared to a chi-squared distribution with degrees of freedom equal for a given tree to the number of values of ω -1, as described in Yang [33]. The phylogenetic data sets supporting the results of this article are available in the TreeBASE repository [http://purl.org/phylo/treebase/phylows/study/TB2:S16421]. The data supporting the evolution rate investigated in this study are included within the article and its additional files (Additional file 1).

### Micro-synteny analyses

Starting from Murat *et al.* [31], chromosomal segments including *MET1* loci were selected from rice and brachypodium. To be able to compare our results with those of Murat *et al.* [31], the same fragment boundaries were retained but in our case, all the coding sequences of a given genomic fragment have been considered. Briefly, rice chromosome 3 [Phytozome: LOC_Os03g58040.1 to LOC_Os03g58920.1] (covering 510.70 kb of genomic DNA

and including 80 genes) and brachypodium chromosome 1 [Phytozome: Bradi1g05680 to Bradi1g04980] (covering 531.9 kb of genomic DNA and including 72 genes) chromosomal segments are syntenic to wheat chromosome 5 and 7 (Additional file 7). Gene sequences from model species were then used to perform BLASTn analysis against the IWGSC sequence surveys as described in Salse et al. [11] using 70% CIP (Cumulative Identity Percentage) but only 30% CALP (Cumulative Alignment Length Percentage). The CALP parameter was kept at a low value in order to detect all the micro-syntenic relationships. Percentage of syntenic conservation was then computed as 100 × the number of conserved genes in wheat/number of genes in the syntenic region from brachypodium or from rice. The data set supporting the results is included within the article and its additional files (Additional file 7).

#### RNA-seq

RNA-seq non-oriented libraries were constructed in two replicates from total RNAs of hexaploid wheat *cv Chinese Spring*. RNAs were prepared with the TruSeq kit (Illumina) for 15 biological samples including 5 organs (root, leaves, stem, spike, grain) and 3 developmental stages (beginning, middle, and end of development) as described in [63] (Additional file 8). For oriented libraries, samples were pooled by organs, rRNAs were removed from total RNAs with the riboZero kit (Ambion) and RNA-seq libraries were constructed with the ScriptSeq kit (Epicentre). All the libraries were sequenced using a HiSeq200 (Illumina) with reads of 100 bp sequenced in both directions. Reads from RNA-seq libraries were mapped using *Tophat2 v2.0.8* [64] and *Bowtie2* [65] onto the *MET1* genomic sequences with neither mismatches nor splice-mismatches allowed. Transcript reconstruction and expression levels were analyzed with *Cufflinks v2.0.2* [66] without annotation. Because sequencing was bidirectional, which is to say that two reads correspond to the same cDNA molecule, expression data results of transcription levels are expressed in Fragments per Kilobase of Exon Model (FPKM) per million mapped reads [67]. The RNA-seq data sets supporting the results of this article are available in the Sequence Read Archive (SRA) repository, [http://www.ncbi.nlm.nih.gov/sra/ERP004714].

#### RNA analyses

Wheat plantlets of *cv Chinese Spring* were grown in a greenhouse and collected at Z61-65 stage according to Zadoks scale [35]. Tissues were frozen in liquid nitrogen and ground to a fine powder. Total RNAs were extracted from 250 mg of plant material using an RNA extraction method adapted from [68]. RNA was subsequently treated with 100 units of DNase I (Invitrogen) in the presence of 20U RNaseOUT™ Recombinant Ribonuclease Inhibitor (Invitrogen). Quantity of extracted RNA was estimated using a Nanodrop (Thermo Scientific) and RNA quality was checked by migration on a 2% agarose gel containing MOPS 2% and 0,05% formaldehyde.

Reverse Transcription was performed from 2 μg of total RNA using an oligo(dT) 15 Primer and M-MLV Reverse Transcriptase (Promega) in presence of Recombinant RNasin Ribonuclease Inhibitor (Promega) according to the supplier's recommendation. Homoeologous specific primers were designed manually and validated with *Oligo Analyzer* (Gene Link) to avoid secondary structure formation. Sequences of selected primer pairs can be found in Additional file 6. Semi-quantitative analyses were performed using primer pairs with similar efficiencies and on the same cDNA sample by comparing the *TaMET1* PCR product to Ta4045 and Ta54227 as reference genes (primer pairs as as in [69]). Quantitative analysis was performed on a LightCycler® 480 System using LightCycler® 480 SYBR Green I Master reagent (Roche) according to the supplier's recommendation. Primer pair efficiencies were calculated through serial dilutions from 1/3 to 1/81 for each RNA sample and only primer pairs with a PCR efficiency between 80 and 110% were selected. As in semi-quantitative analyses, Ta4045 and Ta54227 were used as reference genes.

#### Bisulfite sequencing

1 g of plant material was collected from stem and leaves at the Z30 stage and DNA extracted using the DNeasy plant maxi kit (Qiagen). 200-500 ng of DNA was subjected to bisulfite (BS) treatment using the EZ DNA Methylation-Gold™ Kit (Zymo Research). BS-treated DNA was PCR-amplified using specific primers (Additional file 6) and cloned in pGEMT vectors (Promega) prior to sequencing. 10–20 clones were analyzed for each genomic region using Kismeth software [70]. Two PCR fragments from the VERNALIZATION1 (*VRN1*) gene previously studied by bisulfite experiments [71] were used as internal controls. Incomplete conversion was excluded by analyzing the 0.0 k fragment from *VRN-A1,* which is devoid of CG methylation, while optimal bisulfite treatment were assessed by analysis of the 9.2 k fragment, a highly CG methylated region from *VRN-A1*. Examples of results are given in Additional file 9.

#### Availability of supporting data

The following additional data is available with the online version of this paper. Additional file 1 is a table listing the results of the Likelihood ratio tests. Additional file 2 is a table listing the genetic positions of *TaMET1* loci. Additional file 3 is a sequence alignment of the promoter region of *TaMET1* from homoeologous group 7 with close species. Additional files 4 and 5 are detailed bisulfite analyses performed at *TaMET1* from homoeologous group 5

and 7 respectively. Additional file 6 is a table listing the primers used in this study. Additional file 7 is a table describing micro-synteny data between wheat, rice and brachypodium. Additional file 8 is a table listing the RNA-seq samples used in this study. Additional file 9 is an example of control experiment in bisulfite sequencing analysis.

## Additional files

**Additional file 1: Likelyhood ratio tests (LRT). A)** Likelihood ratio test (LRT). LRT has been used to compare 11 hypotheses ($H_{1-11}$) in respect to the null hypothesis ($H_0$) in which all the branches have the same evolution rate (ω0). Hypotheses were designed to test if the MET1 phylogenetic tree includes up to three evolution rates (ω0, ω1 and ω2) and if these variations in ω values are long term changes (*i.e.* if all the branches below a duplication event display the same ω value) or increase only after a duplication event and then is relaxed to ω0. **B)** Details of the 11 hypotheses tested in the branch model described in Figure 3B. For each hypothesis tested, a tree file in Newick format and a graphic representation highlighting the branches considered in the tested hypothesis are given.

**Additional file 2: Genetic positions of *TaMET1* loci.** Distal and proximal markers from the ITMI reference map and flanking the 2B, 5B and 7A *TaMET1* loci are given in cM.

**Additional file 3: Alignment at putative promoter regions of *TaMET1* genes from homoeologous group 7.** *Hordeum vulgare* chromosome 7 [Ensembl Genomes: MLOC_10988.2], *Triticum aestivum* chromosome 7A [IWGSC: 7AL:4532056], 7B [IWGSC: 7BL:6682174] and 7D [IWGSC: 7DL:3392185], *Triticum urartu* chromosome 7 [Ensembl Genomes: scaffold38640], *Triticum tauschii* chromosome 7 [Ensembl Genomes: scaffold2203], Alignment were performed with *MUSCLE* and refined by *jalview*.

**Additional file 4: Bisulfite analysis of putative promoter region of homoeologous group 5. A)** Frequencies of CG dinucleotides were computed every 50 bp of the putative promoter regions of homoeologous group 5. 5A (black), 5B (white) and 5D (grey). Black bars numbered from 1 to 4 highlight the two regions studied by bisulfite sequencing. **B)** Kismeth outputs of the percentage of methylated cytosines in CG (red), CHG green) and CHH (blue) context.

**Additional file 5: Bisulfite analysis of putative promoter region of homoeologous group 7. A)** Frequencies of CG dinucleotides were computed every 50 bp of the putative promoter regions of homoeologous group 7. 7A (black), 7B (white) and 7D (grey). Black bars numbered from 1 to 4 highlight the two regions studied by bisulfite sequencing. **B)** Kismeth outputs of the percentage of methylated cytosines in CG (red), CHG green) and CHH (blue) context.

**Additional file 6: Primers used in RT-PCR, RT-qPCR, mapping and bisulfite experiments.**

**Additional file 7: Virtual physical map reconstruction at *TaMET1* loci from micro-synteny data.** Physical maps for Os and Bd, virtual physical map based on IWGSC surveys organized from rice and brachypodium orthologs. *TaMET1* loci are highlighted in yellow. Note that two overlapping contigs were found at *TaMET-5A1* indicating that these two IWGSC contigs were not assembled together in the course of the assembly process.

**Additional file 8: RNA-seq samples used in this study.**

**Additional file 9: Controls in bisulfite experiments. A)** Methylation rates at two VRN-A1 regions called 0.0 k and 9.2 k (adapted from [71]). **B)** Structure of the *VRN-A1* gene. **C)** Typical results from bisulfite experiments for 0.0 k (no CG methylation) and 9.2 k (high CG methylation).

### Abbreviations

BAH: Bromo-adjacent homology; BER: Base excision repair; CALP: Cumulative alignment length percentage; CIP: Cumulative identity percentage; CGIs: CpG islands; DNMT1-RFD: Cytosine specific DNA methyltransferase replication foci domain; LRT: Likelihood rate tests; ω: Evolution rate; ESTs: Expressed sequence tags; FPKM: Fragments per kilobase of exon model; gBGC: GC-biaised gene conversion; IWGSC: International wheat genome sequencing consortium; MET1: METHYLTRANSFERASE1; MYa: Million years ago; MITE: Miniature inverted repeat; dN: Rate of non-synonymous substitution; dS: Rate of synonymous substitution; RdDM: RNA-directed DNA Methylation; SNP: Single nucleotide polymorphism; WGD: Whole-genome duplication; Z: Zadoks scale.

### Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

MTho carried out the phylogenetic analyses, SNP mapping, RT-PCR, bisulfite sequencing and participated in sequence capture experiments. EP designed the RNA-seq experiments and LP carried out its analysis. AP designed perl script to compute the frequency of CG dinucleotides. JD carried out the Sequence capture experiments. MThr carried out the syntenic analysis. SF and JPP participated in the design and coordination of the sequence capture, syntenic analysis and genetic mapping. CT designed and coordinated the study and carried out the evolution rate analysis. MTho, AVP and CT wrote the manuscript. All authors read and approved the final manuscript.

### Acknowledgements

MTho is supported by a Cifre PhD agreement n°817/2010 between the CNRS and the Biogemma Company. This work is supported by the "pole de competitivité Céréales Vallée". CTa and AVP are supported by the CNRS, INSERM, Blaise Pascal and Auvergne Universities. AVP is supported by ANR "Dynam'Het"ANR-11 JSV2 009 01 and ANR "SINODYN" ANR-12-ISV6-0001. CTa and AVP are supported by the Region Auvergne through "Life GRID" and a "Young Researcher Fellowship" respectively. We would like to thank Pr D. E. Evans for editing the manuscript, G. Bronner for technical help in evolution rate analysis, J. Enjalbert for sharing results prior to publication, M. Abrouk and F. Choulet for critical reading and helpful suggestions and two anonymous reviewers for their fruitful comments.

### Author details

[1]UMR CNRS 6293 INSERM U 1103 Clermont Université, Genetics Reproduction and Development (GReD), 24 avenue des Landais, BP80026, 63171 Aubière Cedex, France. [2]BIOGEMMA, route d'Ennezat, Centre de Recherche de Chappes, CS 90126, 63720 Chappes, France. [3]UMR INRA 1095 Blaise Pascal University, Genetics Diversity & Ecophysiology of Cereals (GDEC), Clermont-Ferrand – Theix, 5 chemin de Beaulieu, 63039 Clermont-Ferrand Cedex 2, France.

Received: 14 May 2014 Accepted: 13 October 2014
Published: 23 October 2014

### References

1. Gaut BS: **Evolutionary dynamics of grass genomes.** *New Phytol* 2002, **154:**15–28.
2. Feldman M, Lupton F, Miller T: **Wheats.** In *Evol Crops Ed 2 Longman Sci Lond.* Edited by Smartt J, Simmonds N. 1995:184–192.
3. Greer E, Martin AC, Pendle A, Colas I, Jones AME, Moore G, Shaw P: **The Ph1 locus suppresses Cdk2-type activity during premeiosis and meiosis in wheat.** *Plant Cell Online* 2012, **24:**152–162.
4. Moore G, Devos KM, Wang Z, Gale MD: **Cereal genome evolution: grasses, line up and form a circle.** *Curr Biol* 1995, **5:**737–739.
5. Yu J, Hu S, Wang J, Wong GK-S, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J, Chen Y, Huang X, Lin W, Ye C, Tong W, Cong L, Geng J, Han Y, Li L, Li W, Hu G, Huang X, Li W, Li J, Liu Z, Li L, *et al*: **A draft sequence of the rice genome (Oryza sativa L. ssp. indica).** *Science* 2002, **296:**79–92.
6. Goff SA, Ricke D, Lan T-H, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchison D, Martin C, Katagiri F, Lange BM, Moughamer T, Xia Y, Budworth P, Zhong J, Miguel T, Paszkowski U, Zhang S, Colbert M, Sun W, Chen L, Cooper B, Park S, Wood TC, Mao L, Quail P, *et al*: **A draft sequence of the rice genome (Oryza sativa L. ssp. japonica).** *Science* 2002, **296:**92–100.
7. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H,

Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, *et al*: **The Sorghum bicolor genome and the diversification of grasses.** *Nature* 2009, **457:**551–556.

8. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, *et al*: **The B73 maize genome: complexity, diversity, and dynamics.** *Science* 2009, **326:**1112–1115.
9. Vogel JP, Garvin DF, Mockler TC, Schmutz J, Rokhsar D, Bevan MW, Barry K, Lucas S, Harmon-Smith M, Lail K, Tice H, Schmutz (Leader) J, Grimwood J, McKenzie N, Bevan MW, Huo N, Gu YQ, Lazo GR, Anderson OD, Vogel (Leader) JP, You FM, Luo M-C, Dvorak J, Wright J, Febrer M, Bevan MW, Idziak D, Hasterok R, Garvin DF, Lindquist E, *et al*: **Genome sequencing and analysis of the model grass Brachypodium distachyon.** *Nature* 2010, **463:**763–768.
10. Consortium TIBGS: **A physical, genetic and functional sequence assembly of the barley genome.** *Nature* 2012, **491:**711–716.
11. Salse J, Bolot S, Throude M, Jouffe V, Piegu B, Quraishi UM, Calcagno T, Cooke R, Delseny M, Feuillet C: **Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution.** *Plant Cell Online* 2008, **20:**11–24.
12. Akhunov ED, Akhunova AR, Linkiewicz AM, Dubcovsky J, Hummel D, Lazo G, Chao S, Anderson OD, David J, Qi L, Echalier B, Gill BS, Miftahudin, Gustafson JP, Rota ML, Sorrells ME, Zhang D, Nguyen HT, Kalavacharla V, Hossain K, Kianian SF, Peng J, Lapitan NLV, Wennerlind EJ, Nduati V, Anderson JA, Sidhu D, Gill KS, McGuire PE, Qualset CO, *et al*: **Synteny perturbations between wheat homoeologous chromosomes caused by locus duplications and deletions correlate with recombination rates.** *Proc Natl Acad Sci* 2003, **100:**10836–10841.
13. Choulet F, Wicker T, Rustenholz C, Paux E, Salse J, Leroy P, Schlub S, Le Paslier M-C, Magdelenat G, Gonthier C, Couloux A, Budak H, Breen J, Pumphrey M, Liu S, Kong X, Jia J, Gut M, Brunel D, Anderson JA, Gill BS, Appels R, Keller B, Feuillet C: **Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces.** *Plant Cell Online* 2010, **22:**1686–1701.
14. Kellis M, Birren BW, Lander ES: **Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae.** *Nature* 2004, **428:**617–624.
15. Smet RD, Adams KL, Vandepoele K, Montagu MCEV, Maere S, Peer YV d: **Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants.** *Proc Natl Acad Sci* 2013, **110:**2898–2903.
16. Schnable JC, Springer NM, Freeling M: **Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss.** *Proc Natl Acad Sci* 2011, **108:**4069–4074.
17. Cheng F, Wu J, Fang L, Sun S, Liu B, Lin K, Bonnema G, Wang X: **Biased gene fractionation and dominant gene expression among the subgenomes of brassica rapa.** *PLoS One* 2012, **7:**e36442.
18. Mayer KFX, Rogers J, Doležel J, Pozniak C, Eversole K, Feuillet C, Gill B, Friebe B, Lukaszewski AJ, Sourdille P, Endo TR, Kubaláková M, Číhalíková J, Dubská Z, Vrána J, Šperková R, Šimková H, Febrer M, Clissold L, McLay K, Singh K, Chhuneja P, Singh NK, Khurana J, Akhunov E, Choulet F, Alberti A, Barbe V, Wincker P, Kanamori H, *et al*: **A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome.** *Science* 2014, **345:**1251788.
19. Groszmann M, Greaves IK, Fujimoto R, James Peacock W, Dennis ES: **The role of epigenetics in hybrid vigour.** *Trends Genet* 2013, **29:**684–690.
20. Jackson S, Chen ZJ: **Genomic and expression plasticity of polyploidy.** *Curr Opin Plant Biol* 2010, **13:**153–159.
21. Feng S, Cokus SJ, Zhang X, Chen P-Y, Bostick M, Goll MG, Hetzel J, Jain J, Strauss SH, Halpern ME, Ukomadu C, Sadler KC, Pradhan S, Pellegrini M, Jacobsen SE: **Conservation and divergence of methylation patterning in plants and animals.** *Proc Natl Acad Sci U S A* 2010, **107:**8689–8694.
22. Mirouze M, Reinders J, Bucher E, Nishimura T, Schneeberger K, Ossowski S, Cao J, Weigel D, Paszkowski J, Mathieu O: **Selective epigenetic control of retrotransposition in Arabidopsis.** *Nature* 2009, **461:**427–430.
23. Lippman Z, Gendrel A-V, Black M, Vaughn MW, Dedhia N, McCombie WR, Lavine K, Mittal V, May B, Kasschau KD, Carrington JC, Doerge RW, Colot V, Martienssen R: **Role of transposable elements in heterochromatin and epigenetic control.** *Nature* 2004, **430:**471–476.
24. Kinoshita Y, Saze H, Kinoshita T, Miura A, Soppe WJJ, Koornneef M, Kakutani T: **Control of FWA gene silencing in Arabidopsis thaliana by SINE-related direct repeats.** *Plant J* 2006, **49:**38–45.
25. Melamed-Bessudo C, Levy AA: **Deficiency in DNA methylation increases meiotic crossover rates in euchromatic but not in heterochromatic regions in Arabidopsis.** *Proc Natl Acad Sci* 2012, **109:**E981–E988.
26. Vongs A, Kakutani T, Martienssen RA, Richards EJ: **Arabidopsis thaliana DNA methylation mutants.** *Science* 1993, **260:**1926–1928.
27. Steward N, Kusano T, Sano H: **Expression of ZmMET1, a gene encoding a DNA methyltransferase from maize, is associated not only with DNA replication in actively proliferating cells, but also with altered DNA methylation status in cold-stressed quiescent cells.** *Nucleic Acids Res* 2000, **28:**3250–3259.
28. Teerawanichpan P, Chandrasekharan M, Jiang Y, Narangajavana J, Hall T: **Characterization of two rice DNA methyltransferase genes and RNAi-mediated reactivation of a silenced transgene in rice callus.** *Planta* 2004, **218:**337–349.
29. Fujimoto R, Sasaki T, Nishio T: **Characterization of DNA methyltransferase genes in *Brassica rapa*.** *Genes Genet Syst* 2006, **81:**235–242.
30. Brenchley R, Spannagl M, Pfeifer M, Barker GLA, D'Amore R, Allen AM, McKenzie N, Kramer M, Kerhornou A, Bolser D, Kay S, Waite D, Trick M, Bancroft I, Gu Y, Huo N, Luo M-C, Sehgal S, Gill B, Kianian S, Anderson O, Kersey P, Dvorak J, McCombie WR, Hall A, Mayer KFX, Edwards KJ, Bevan MW, Hall N: **Analysis of the bread wheat genome using whole-genome shotgun sequencing.** *Nature* 2012, **491:**705–710.
31. Murat F, Xu J-H, Tannier E, Abrouk M, Guilhot N, Pont C, Messing J, Salse J: **Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution.** *Genome Res* 2010, **20:**1545–1557.
32. Warren AS, Anandakrishnan R, Zhang L: **Functional bias in molecular evolution rate of Arabidopsis thaliana.** *BMC Evol Biol* 2010, **10:**125.
33. Yang Z: **PAML 4: Phylogenetic Analysis by Maximum Likelihood.** *Mol Biol Evol* 2007, **24:**1586–1591.
34. Duret L, Mouchiroud D: **Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate.** *Mol Biol Evol* 2000, **17:**68–070.
35. Zadoks JC, Chang TT, Konzak CF: **A decimal code for the growth stages of cereals.** *Weed Res* 1974, **14:**415–421.
36. Jullien PE, Susaki D, Yelagandula R, Higashiyama T, Berger F: **DNA Methylation dynamics during sexual reproduction in Arabidopsis thaliana.** *Curr Biol* 2012, **22:**1825–1830.
37. Yamauchi T, Moritoh S, Johzuka-Hisatomi Y, Ono A, Terada R, Nakamura I, Iida S: **Alternative splicing of the rice OsMET1 genes encoding maintenance DNA methyltransferase.** *J Plant Physiol* 2008, **165:**1774–1782.
38. Choulet F, Alberti A, Theil S, Glover N, Barbe V, Daron J, Pingault L, Sourdille P, Couloux A, Paux E, Leroy P, Mangenot S, Guilhot N, Gouis JL, Balfourier F, Alaux M, Jamilloux V, Poulain J, Durand C, Bellec A, Gaspin C, Safar J, Dolezel J, Rogers J, Vandepoele K, Aury J-M, Mayer K, Berges H, Quesneville H, Wincker P, *et al*: **Structural and functional partitioning of bread wheat chromosome 3B.** *Science* 2014, **345:**1249721.
39. Genger RK, Kovac KA, Dennis ES, Peacock WJ, Finnegan EJ: **Multiple DNA methyltransferase genes in Arabidopsis thaliana.** *Plant Mol Biol* 1999, **41:**269–278.
40. Finnegan EJ, Kovac KA: **Plant DNA methyltransferases.** *Plant Mol Biol* 2000, **43:**189–201.
41. Chen M, Ha M, Lackey E, Wang J, Chen ZJ: **RNAi of met1 reduces DNA methylation and induces genome-specific changes in gene expression and centromeric small RNA accumulation in Arabidopsis Allopolyploids.** *Genetics* 2008, **178:**1845–1858.
42. Diez CM, Roessler K, Gaut BS: **Epigenetics and plant genome evolution.** *Curr Opin Plant Biol* 2014, **18:**1–8.
43. Tran RK, Henikoff JG, Zilberman D, Ditt RF, Jacobsen SE, Henikoff S: **DNA Methylation profiling identifies CG methylation clusters in Arabidopsis Genes.** *Curr Biol* 2005, **15:**154–159.
44. Lynch M, Conery JS: **The evolutionary fate and consequences of duplicate genes.** *Science* 2000, **290:**1151–1155.
45. Yang L, Gaut BS: **Factors that contribute to variation in evolutionary rate among Arabidopsis Genes.** *Mol Biol Evol* 2011, **28:**2359–2369.
46. Yang L, Takuno S, Waters ER, Gaut BS: **Lowly expressed genes in Arabidopsis thaliana bear the signature of possible pseudogenization by promoter degradation.** *Mol Biol Evol* 2011, **28:**1193–1203.

47. Duret L, Galtier N: **Biased gene conversion and the evolution of mammalian genomic landscapes.** *Annu Rev Genomics Hum Genet* 2009, **10:**285–311.
48. Smith ZD, Meissner A: **DNA methylation: roles in mammalian development.** *Nat Rev Genet* 2013, **14:**204–220.
49. Ponger L, Duret L, Mouchiroud D: **Determinants of CpG islands: expression in early embryo and isochore structure.** *Genome Res* 2001, **11:**1854–1860.
50. Suzuki S, Shaw G, Kaneko-Ishino T, Ishino F, Renfree MB: **The evolution of mammalian genomic imprinting was accompanied by the acquisition of novel CpG islands.** *Genome Biol Evol* 2011, **3:**1276–1283.
51. Serres-Giardi L, Belkhir K, David J, Glémin S: **Patterns and evolution of nucleotide landscapes in seed plants.** *Plant Cell Online* 2012, **24:**1379–1397.
52. Saintenac C, Falque M, Martin OC, Paux E, Feuillet C, Sourdille P: **Detailed recombination studies along chromosome 3B provide new insights on crossover distribution in wheat (Triticum aestivum L.).** *Genetics* 2009, **181:**393–403.
53. Yamauchi T, Johzuka-Hisatomi Y, Terada R, Nakamura I, Iida S: **The MET1b gene encoding a maintenance DNA methyltransferase is indispensable for normal development in rice.** *Plant Mol Biol* 2014, **85:**219–232.
54. Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, Richmond TA, Middle CM, Rodesch MJ, Packard CJ, Weinstock GM, Gibbs RA: **Direct selection of human genomic loci by microarray hybridization.** *Nat Methods* 2007, **4:**903–905.
55. Saintenac C, Jiang D, Akhunov ED: **Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome.** *Genome Biol* 2011, **12:**R88.
56. Gervais L, Dedryver F, Morlais J-Y, Bodusseau V, Negre S, Bilous M, Groos C, Trottet M: **Mapping of quantitative trait loci for field resistance to Fusarium head blight in an European winter wheat.** *Theor Appl Genet* 2003, **106:**961–970.
57. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32:**1792–1797.
58. Castresana J: **Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis.** *Mol Biol Evol* 2000, **17:**540–552.
59. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O: **New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0.** *Syst Biol* 2010, **59:**307–321.
60. Letunic I, Bork P: **Interactive tree of life v2: online annotation and display of phylogenetic trees made easy.** *Nucleic Acids Res* 2011, **39**(suppl 2): W475–W478.
61. Salamov AA, Solovyev VV: **Ab initio gene finding in Drosophila Genomic DNA.** *Genome Res* 2000, **10:**516–522.
62. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite.** *Trends Genet* 2000, **16:**276–277.
63. Rustenholz C, Choulet F, Laugier C, Šafář J, Šimková H, Doležel J, Magni F, Scalabrin S, Cattonaro F, Vautrin S, Bellec A, Bergès H, Feuillet C, Paux E: **A 3,000-loci transcription map of chromosome 3B unravels the structural and functional features of gene islands in Hexaploid Wheat.** *Plant Physiol* 2011, **157:**1596–1608.
64. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq.** *Bioinformatics* 2009, **25:**1105–1111.
65. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10:**R25.
66. Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L: **Improving RNA-Seq expression estimates by correcting for fragment bias.** *Genome Biol* 2011, **12:**R22.
67. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5:**621–628.
68. Bogorad L, Gubbins EJ, Krebbers ET, Larrinua IM, Mulligan BJ, Muskavitch KMT, Orr EA, Rodermel SR, Schantz R, Steinmetz AA, De Vos G, Ye YK: **Cloning and physical mapping of maize plastid genes.** *Methods Enzymol* 1983, **97:**524–554.
69. Paolacci AR, Tanzarella OA, Porceddu E, Ciaffi M: **Identification and validation of reference genes for quantitative RT-PCR normalization in wheat.** *BMC Mol Biol* 2009, **10:**11.
70. Gruntman E, Qi Y, Slotkin RK, Roeder T, Martienssen RA, Sachidanandam R: **Kismeth: analyzer of plant methylation states through bisulfite sequencing.** *BMC Bioinformatics* 2008, **9:**371–371.
71. Khan A, Enjalbert J, Marsollier A-C, Rousselet A, Goldringer I, Vitte C: **Vernalization treatment induces site-specific DNA hypermethylation at the VERNALIZATION-A1 (VRN-A1) locus in hexaploid winter wheat.** *BMC Plant Biol* 2013, **13:**209.

doi:10.1186/1471-2164-15-922
**Cite this article as:** Thomas *et al.*: Evolutionary history of Methyltransferase 1 genes in hexaploid wheat. *BMC Genomics* 2014 **15**:922.